DOCUMENT RESUME

ED 093 342 95 IR 000 829

AUTHOR Ajay, Helen B.; And Others
TITLE Analysis of Essays by Computer (AEC-II). Final Report.
INSTITUTION Connecticut Univ., Storrs. Dept. of Psychology.
SPONS AGENCY National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
BUREAU NO BR-8-0102
PUB DATE Dec 73
GRANT OEG-0-9-160108-3712(010)
NOTE 289p.

EDRS PRICE MF-$0.75 HC-$13.80 PLUS POSTAGE
DESCRIPTORS *Computer Programs; *Essay Tests; Feasibility Studies; Predictive Ability (Testing); Program Evaluation; *Scoring; *Secondary School Students

ABSTRACT

After years of successful experimental grading of student essays by computer, it was necessary to design and test certain meta-strategies for automatic use in American secondary schools. First, schools were divided into four subject-matter fields (science, social studies, English, and foreign language) and three grade-levels (7-8, 9-10, 11-12). Two classes were recruited in each cell, making 24 in all, and two essaytest items, written by the teacher, provided large and diverse data sets. The student responses were independently rated by four competent human judges, and the goal of the research became the optimal simulation of these ratings. Computer keys were generated in two ways: supplied by the teacher; and adopted from a "model" student essay chosen from the class responses. A combination of strategies, some general and stylistic, and some specific and key-based, were tested for each item. By combining strategies across items and subject-fields and grade-levels, large general principles were discovered, suggesting eventual practical use across a wide range of classroom settings. (Author)

ED 093342

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Final Report

Project No. 8-0102
Grant No. OEG-0-9-160108-3712(010)

BEST COPY AVAILABLE

ANALYSIS OF ESSAYS BY COMPUTER (AEC-II)

Helen B. Ajay, Peter I. Tillett, and Ellis B. Page

Ellis B. Page, Principal Investigator
Department of Educational Psychology
University of Connecticut
Storrs, Connecticut 06268

December 1973

IR 000 829

The research reported herein was performed pursuant to a grant with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education

BEST COPY AVAILABLE

National Center for Educational Research and Development

This document was processed for the ERIC Document Reproduction Service by the ERIC Clearinghouse at Stanford. We are aware that some pages probably will not be readable in microfiche or in Hard copy form. However, this is the best available copy, and we feel that the document should not be withheld from interested readers on the basis of these unreadable pages alone

ANALYSIS OF ESSAYS BY COMPUTER (AEC-II)

Helen B. Ajay, Peter I. Tillett, and Ellis B. Page

(Abstract)

After years of successful experimental grading of student essays by computer, it was necessary to design and test certain meta-strategies for automatic use in American secondary schools. First, schools were divided into four subject-matter fields (science, social studies, English, and foreign language) and three grade-levels (7-8, 9-10, 11-12). Two classes were recruited in each cell, making 24 in all, and two essay-test items, written by the teacher, provided large and diverse data sets. The student responses were independently rated by four competent human judges, and the goal of the research became the optimal simulation of these ratings. Computer keys were generated in two ways: supplied by the teacher; and adopted from a "model" student essay chosen from the class responses. A combination of strategies, some general and stylistic, and some specific and key-based, were tested for each item. By combining strategies across items and subject-fields and grade-levels, large general principles were discovered, suggesting eventual practical use across a wide range of classroom settings.

ACKNOWLEDGEMENTS

The authors of this report have debts to many people, who have contributed to the work from its inception, late in 1964. In the earlier work on stylistic essay-grading, the authors received good advice from Junias A. Davis, Allan B. Ellis, and many others. That work was supported, first, by the College Entrance Examination Board, and second, by the U. S. Office of Education (AEC-I). Gerald and Mary Ann Fisher were early programmers of the first analysis, and inspired some of the later computer work as well. Dieter Paulus was an important contributor and co-investigator of that earlier work.

Contributory work was also done by Donald R. Marcotte and John McManus (who also commented on this manuscript in an earlier form). Drs. Marcotte and McManus, and Thomas F. Breen III did work in prose analysis which provided some of the insights necessary to the present design.

Valuable comments on the current work were also received from Robert McMorris, Jack Hiller, Dieter Paulus, William McMunn, Richard Otte, and Daniel Antonopoulos. Great support was given by the University Research Foundation at the University of Connecticut, led by Dean Hugh Clark, and by the University Computer Center. Miss Louise Patros was of invaluable assistance in administering the earlier work and a portion of this current one. Special mention should be made of the contribution of Dr. Christopher Longyear, Research Associate and Computational Linguist for the project in its first two years, 1969-71. And John B. Carroll gave excellent advice.

At this writing, work with the computer analysis of student prose is about nine years old. A large number of people have engaged in the work, have verified the earlier findings, and have broadened the scope. Particular mention should be made of Henry Slotnick, who has advanced theory of stylistic analysis, and of Thomas Knapp, who has illuminated some statistical strategies. But surely, many names which should be included in this acknowledgement, will be neglected for lack of space.

Some word should be said about the roles played in this report. E. B. Page was the Principal Investigator, and played the largest single role in the research design, as he had in that of the earlier work (AEC-I) in stylistic analysis. Dr. Peter I. Tillett (now of the South Australia Board of Education) organized the large computer system which accomplished the analysis (certain principles and insights were borrowed from the work by Drs. Longyear, Fisher, and Marcotte). And Dr. Helen B. Ajay (now of the Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, California) performed much of the statistical analysis and interpretation of the data. Her doctoral dissertation served as the summary document for the most important findings of the project, and has been incorporated as such a summary in this Report. These roles account for the authorship as listed.

E.B.P., Storrs, Connecticut

TABLE OF CONTENTS

		Page
LIST OF TABLES		vi
LIST OF FIGURES		xi
CHAPTER		
I	STATEMENT OF THE PROBLEM	1
	Background of the Problem	1
	Statement of the Problem	3
	Statement of Hypotheses	5
	Significance of the Study for Education	7
	Summary	9
II	BACKGROUND LITERATURE	10
	Introduction	10
	The Page and Paulus Study	11
	McManus Study	12
	Marcotte Study	14
	Content Analyses Using "Vagueness Concept"	16
	Whalen Study	17
	Archambault and Greene Studies	19
	Summary	20
III	PROCEDURES	21
	Data Collection Procedures	21
	Rationale	21
	Essay Handling Procedures	24
	Essay Rating Procedures	27
	Computer Strategies for Content Analysis	28
	Introduction	28
	Stage 1	31
	Stage 2	34
	Manual Procedures	34
	Stage 2 Computer Program Description	37
	Stage 3, Run 1	38
	Stage 3, Run 2	41
	Stage 3, Supplementary Run 2A	41
	Stage 3, Run 3	42
	The "Standard Dictionary"	43
	The Proper Noun Dictionary	43
	The Spelling Error Dictionary	43

Choice of Root-Words..... 44
The Stemming Algorithm..... 45
Stage 4..... 46
Hypotheses-Testing Procedures..... 49
Content Specificity..... 52
Summary..... 53

IV RESULTS..... 55
Descriptive Measures..... 56
Judges' Ratings of the Essays..... 56
Judges' Ratings of Teacher Key and Model Student Essay Effectiveness..... 63
Judges; Ratings of Essay Question "Content Specificity..... 68
Multiple Regression Analyses..... 76
Introduction..... 76
Analyses Using All Essays..... 80
Analyses By Subject Area..... 101
Analyses by Grade Level..... 128
Analyses by Content Specificity Level..... 145
Summary..... 167

V DISCUSSION AND IMPLICATIONS FOR FUTURE RESEARCH.... 170
Discussion..... 170
Judge Reliability..... 170
Analyses of All Essays..... 171
Analyses by Subject Area..... 172
Analyses by Grade Level..... 173
Analyses by Content Specificity Level..... 175
Overall Success of the Independent Variables in Prediction..... 177
Major Hypothesis-Testing Results..... 184
Implications for Future Research..... 187
Computer Algorithms..... 187
The "Thesaurus Problem"..... 189
Content Specificity..... 190
Content-Related Predictors..... 192
Restricted Models..... 193
Summary..... 193

BIBLIOGRAPHY..... 195

APPENDICES..... 205

A	Instructions to Essay Raters.....	205
B	List of Function Words Used As Input to Stage 3, Run 1.....	208
C	The Stemming Algorithm (Stage 3, Run 3).....	210
~~D~~	~~Output From Stages 2 Through 4.....~~	~~214~~
~~E~~	Instructions to Raters, Ratings of Essay Questions as to Content Specificity and Emphasis.....	229

LIST OF TABLES

TABLE		Page
3.1	Coding System for Identification of AEC Essays	26
3.2	Thesaurie Categories and Number of Synonyms Appearing in Roget's Thesaurus for the Word "Land"	35
4.1	Judges' Ratings, Means and Standard Deviations, Science Essays	57
4.2	Judges' Ratings, Means and Standard Deviations, Social Studies Essays	58
4.3	Judges' Ratings, Means and Standard Deviations, English Essays	59
4.4	Judges' Ratings, Means and Standard Deviations, Foreign Language Essays	60
4.5	Intercorrelations of Essay Ratings for Judge Pairs (By Subject Area)	61
4.6	Pooled Reliability Estimates for All Judges and All Combinations of Three Judges Using Analysis of Variance (By Subject Area)	62
4.7	Means and Standard Deviations, Teacher Key Evaluations, Before and After Rating Student Essays (Four-Point Rating Scale)	64
4.8	Means and Standard Deviations, Model Student Essay Evaluations, Before and After Rating Student Essays (Four-Point Rating Scale)	66
4.9	Intercorrelations of Judges' Evaluations of Teacher Key Effectiveness Before and After Rating Student Essays	67

4.10 Intercorrelations of Judges' Evaluations of Model Student Essay Effectiveness Before and After Rating Student Essays.....69

4.11 Correlations Between "Before" and "After" Ratings, Teacher Keys and Model Student Essays.....70

4.12 Means and Standard Deviations, Content Specificity Ratings (Ten-Point Scale).....72

4.13 Judge Intercorrelations, Content Specificity Ratings..73

4.14 Pooled Reliability Estimates for All Judges and All Combinations of Four Judges Using Analysis of Variance, Content Specificity Ratings.....75

4.15 Intercorrelation Matrix, All Essays, Stylistic and Content Variables (Essay as Unit).....81

4.16 Intercorrelation Matrix, All Essays, Stylistic and Content Variables (Sentence as Unit).....82

4.17 Stepwise Multiple Regression, All Essays, Stylistic Variables Only.....84

4.18 Stepwise Multiple Regression, All Essays, Stylistic and Content Variables (Essay as Unit).....89

4.19 Stepwise Multiple Regression, All Essays, Stylistic and Content Variables (Sentence as Unit).....91

4.20 Stepwise Multiple Regression, All Essays, Stylistic and Combined Content Variables (Essay as Unit).....93

4.21 Stepwise Multiple Regression, All Essays, Stylistic and Combined Content Variables (Sentence as Unit).....95

4.22 Grouping of Student Essays for Cross-Validation Procedure.....97

4.23 Cross-Validation Results.....98

4.24 Summary of Analyses Using All Essays.....100

4.25 Intercorrelation Matrix, Science Essays, Stylistic and Content Variables.....102

4.26 Intercorrelation Matrix, Social Studies Essays, Stylistic and Content Variables.....103

4.27 Intercorrelation Matrix, English Essays, Stylistic and Content Variables.....104

4.28 Intercorrelation Matrix, Foreign Language Essays, Stylistic and Content Variables.....105

4.29 Stepwise Multiple Regression, Science Essays, Stylistic Variables Only.....108

4.30 Stepwise Multiple Regression, Social Studies Essays, Stylistic Variables Only.....110

4.31 Stepwise Multiple Regression, English Essays, Stylistic Variables Only.....111

4.32 Stepwise Multiple Regression, Foreign Language Essays, Stylistic Variables Only.....113

4.33 Stepwise Multiple Regression, Science Essays, Content and Stylistic Variables (Essay as Unit).....115

4.34 Stepwise Multiple Regression, Social Studies Essays, Content and Stylistic Variables (Essay as Unit).....116

4.35 Stepwise Multiple Regression, English Essays, Content and Stylistic Variables (Essay as Unit).....118

4.36 Stepwise Multiple Regression, Foreign Language Essays, Content and Stylistic Variables (Essay as Unit).....119

4.37 Stepwise Multiple Regression, Science Essays, Stylistic and Content Variables (Sentence as Unit)...121

4.38 Stepwise Multiple Regression, Social Studies Essays, Stylistic and Content Variables (Sentence as Unit)...122

4.39 Stepwise Multiple Regression, English Essays, Stylistic and Content Variables (Sentence as Unit).....124

4.40 Stepwise Multiple Regression, Foreign Language Essays, Stylistic and Content Variables (Sentence as Unit).....125

4.41 Summary of Analyses by Subject Area.....127

4.42 Stepwise Multiple Regression, Seventh and Eighth Grade Essays, Stylistic Variables Only.....129

4.43 Stepwise Multiple Regression, Ninth and Tenth Grade Essays, Stylistic Variables Only.....130

4.44 Stepwise Multiple Regression, Eleventh and Twelfth Grade Essays, Stylistic Variables Only.....131

4.45 Stepwise Multiple Regression, Seventh and Eighth Grade Essays, Stylistic and Content Variables (Essay as Unit).....133

4.46 Stepwise Multiple Regression, Ninth and Tenth Grade Essays, Stylistic and Content Variables (Essay as Unit).....134

4.47 Stepwise Multiple Regression, Eleventh and Twelfth Grade Essays, Stylistic and Content Variables (Essay as Unit).....136

4.48 Stepwise Multiple Regression, Seventh and Eighth Grade Essays, Stylistic and Content Variables (Sentence as Unit).....137

4.49 Stepwise Multiple Regression, Ninth and Tenth Grade Essays, Stylistic and Content Variables (Sentence as Unit).....139

4.50 Stepwise Multiple Regression, Eleventh and Twelfth Grade Essays, Stylistic and Content Variables (Sentence as Unit).....141

4.51 Summary of Analyses by Grade Level.....142

4.52 Grouping of Essay Questions by Content Specificity...146

4.53 Stepwise Multiple Regression, Content Specificity--High Group, Stylistic Variables Only.....148

4.54 Stepwise Multiple Regression, Content Specificity--Middle Group, Stylistic Variables Only.....149

4.55 Stepwise Multiple Regression, Content Specificity--Low Group, Stylistic Variables Only.....150

4.56 Stepwise Multiple Regression, Content Specificity--High Group, Stylistic and Content Variables (Essay as Unit).....152

4.57 Stepwise Multiple Regression, Content Specificity--Middle Group, Stylistic and Content Variables (Essay as Unit).....154

4.58 Stepwise Multiple Regression, Content Specificity--Low Group, Stylistic and Content Variables (Essay as Unit).....156

4.59 Stepwise Multiple Regression, Content Specificity--High Group, Stylistic and Content Variables (Sentence as Unit).....157

4.60 Stepwise Multiple Regression, Content Specificity--Middle Group, Stylistic and Content Variables (Sentence as Unit).....158

4.61 Stepwise Multiple Regression, Content Specificity--Low Group, Stylistic and Content Variables (Sentence as Unit).....160

4.62 Stepwise Multiple Regression, Content Specificity--All Groups, Content Variables Only (Essay as Unit)...161

4.63 Stepwise Multiple Regression, Content Specificity--All Groups, Content Variables Only (Sentence as Unit).....163

4.64 Summary of Analyses by Content Specificity Level.....165

LIST OF FIGURES

Figure		Page
3.1	Data Collection for the AEC-II Project.....	22
3.2	General Strategy for Content Analysis.....	30
3.3	Runs Involved in Stage 3.....	39

CHAPTER I

STATEMENT OF THE PROBLEM

Background of the Problem

The use of the essay examination as an instrument for evaluating scholarly performance in a subject area continues after more than a half-century-long effort to introduce objective testing. One reason for this phenomenon may be the persistent belief among educators that the essay examination provides, in a way that the objective test does not, a means of assessing the student's ability to assemble, organize, and communicate both his ideas on a topic and the evidence to support a position regarding these ideas. Implied in this judgment is the following criticism of objective tests: "...it isn't enough to be able to recognize a correct fact when it is presented or to discriminate among alternatives posed by others; the educated man is the master of his collection of ideas" (Coffman, 1971, p. 273).

Even if one accepts this judgment of objective tests, however, there are a number of telling factors which argue for the use of objective, multiple-choice-type tests over the use of essay examinations. These arguments mainly concern the practical problems involved in the administration and evaluation of essay tests. As

Page (1969, p. 2) has pointed out:

> A single judgment of an essay by a single human judge is slow, extremely unreliable, and of uncertain status. When sufficient training is used, and a sufficient number of judgments establish a decent reliability, essay analysis becomes prohibitively expensive. For measurement purposes, therefore, psychometricians have settled for multiple-choice items. These have the virtues of wide sampling, since more questions may be asked within a given time period; of high reliability; and of defensive validity, since scores often correlate as highly with judgmental ratings as the ratings correlate with each other under ordinary conditions.

Research into the feasibility of devising adequate models and techniques for the computer analysis of essays was begun over five years ago in an attempt to overcome some of the time limitations and unreliability problems associated with essay evaluation; to learn more about the nature and structure of the essay question and response and the problems encountered in both the taking and the grading of essay examinations; and, ultimately, to provide the teacher with a reasonable tool for the evaluation of essays and for diagnostic purposes, leaving the teacher with more time for direct contact with the students. Major contributions to this area of research have been made by faculty and students of the School of Education at the University of Connecticut. The present study has developed from this work.

Statement of the Problem

An essay response is by definition more difficult to evaluate than responses to objective test items, since the essay examination is characterized by "the freedom of responses allowed the examinee and the fact that not only can no single answer be listed as correct and complete, and given to clerks to check, but even an expert cannot usually classify a response as categorically right or wrong. Rather, there are different degrees of quality or merit which can be recognized" (Stallnaker, 1951, p. 495). While certain extensive responses categorized as essays (e.g., translations of foreign language passages) may fit into a more rigorous mold, allowing less freedom of response on the part of the student, the essay grader usually faces a rather complex problem, his judgment resulting from his perceptions regarding the interactions of a number of variables. A variety of qualities may be sought in the essay response which in an objective test response are to be found in a very reduced (specific) form or are entirely lacking. Regarding the qualities an essay may possess, the distinction between its "style" and its "content" is the most elementary breakdown. The "Analysis of Essays by Computer" (AEC-II) Project currently being undertaken at the University of Connecticut is concerned with the investigation of both essay content and style, and

is a logical sequel to a previous study (Page and Paulus, 1968) in which a number of stylistic features of essay responses were analyzed and used as predictors of essay grades.

In their 1968 study, Page and Paulus presented a rationale for the content analysis of student essays, utilizing the following procedure:

> The employment of a simple lexicon of key words, to the acceptance of their synonyms, to the search for the key words or synonyms in appropriate contexts, to the search for these meanings in appropriate relationships. (p. 193)

The present study, "Strategies for the Content Analysis of Essays by Computer," has been designed to comprise a portion of the AEC-II Project and utilizes a model based on the rationale proposed by Page and Paulus in 1968. Given an initial general model for the content analysis of essays by computer, the main tasks of the study were: (1) to develop "computer-usable keys" for the essay questions in the available data set (these keys to be developed on the basis of the key words and contingencies found in the teacher key for each essay question and the "model student essay" -- best student response as determined by the teacher devising the test -- for each question) using the scheme dictated by the model and compatible with the computer algorithms devised for applying the model, and (2) to determine the efficiency of the model by apply-

ing the strategy (i.e., the keys and the algorithms) to the essays and comparing the "variable scores" produced by "matches" using this method with the scores given the essays by human judges. Degrees of efficiency of the model were compared across subject area, grade level, and "content specificity" dimensions.

Statement of Hypotheses

The major hypothesis underlying the present study was that the inclusion of variables derived from considerations of both content and style in a multiple regression analysis of student essays would prove more effective in predicting student grades (i.e., yield a significantly higher MULTR value) than would the use of stylistic variables alone. The criterion measure, "essay grade," is defined here as the average grade assigned to an essay by four "judges" (teachers in the subject area represented by an essay question). It was expected that the significantly greater predictive value derived from the use of both content and stylistic variables in the multiple regression equation would hold across: (1) the four subject areas under investigation, (2) the three grade levels represented by the data, (3) the three "content specificity" levels (defined below) identified in the data.

Further, it was hypothesized that gain in the MULTR values from the addition of content variables to the prediction equation

would be greater for essays in the two subject areas (science and foreign language) in which the items appeared to call for relatively more specific responses than did the items in the other two subject areas (English and social studies).

Three grade levels were represented in the data: 7th-8th, 9th-10th, and 11th-12th. A third hypothesis was that gain in MULTR values resulting from the addition of content variables to the equation would be greater for the 7th-8th grade level than for the 11th-12th grade level, under the assumption that in the lower grade level, content, more than stylistic considerations, would be emphasized.

"Content specificity" was defined as the degree to which an essay item emphasized content rather than style in its requirements for a response. Each of the 48 essay items in the data were rated for degree of content specificity (ten-point scale) by five judges; the items were then rank-ordered on the basis of the average rating given to each. This procedure enabled the division of the essays associated with the items into three groups representing "high," "middle," and "low" content specificity. Regarding this quality, it was hypothesized that gain in MULTR value resulting from the addition of content predictors to the multiple regression equation would be greater for the high content specificity group

than for the low content specificity group in the data.

Finally, two different "student response units" (or "search units") were used to determine values for the content-related predictors -- (1) the whole student essay, and (2) a single sentence within the student essay. It was hypothesized that equally good prediction would result when either the essay as a whole was used as the search unit for matches on the content variables, or when the sentence was used as the search unit for matches on the content variables.

Significance of the Study for Education

Carroll, in his Study of Language (1953, p. 6), wrote:

> ...the study of language and communication phenomena... is beginning to emerge as one of the most fundamental disciplines in the psychological and social sciences -- quite as important, for example, as the study of molecular physics in the natural and biological sciences.

The present study and its predecessors lie in the mainstream of this important and relatively new branch of empirical research.

The objective test, as indicated in the opening paragraphs of this chapter, almost always lacks certain characteristics which educators find highly desirable in evaluation instruments, most particularly in those disciplines where freedom of expression on the part of the student is necessary for determining the student's ability and achievement in the discipline (e.g., English studies).

Also, there is evidence to support the contention that the use of objective tests creates highly specific study patterns in students (see Merkhofer, 1954). Meyer (1934, 1935) found that greater achievement on any type of examination followed study in anticipation of an essay examination. In addition, good objective tests (those which will adequately cover the subject matter and measure complex skills and understandings), while very easy to score, are extremely difficult to prepare. Coffman (1972, p. 1) describes the typical classroom teacher's frustration in this regard and hypothesizes that this is one major reason that the teacher is "likely to return to the familiar essay examination whenever he wishes to do more than measure the student's knowledge of facts".

Ratings obtained from judges on essay examinations, on the other hand, suffer greatly from lack of both objectivity and reliability. In an extensive discussion of this topic, Coffman (1971, p. 296) indicates that work on the computer grading of essays holds promise for injecting greater reliability into essay ratings, and specifically cites the pioneering work done in this area at the University of Connecticut by Page et al. In his proposal for the current AEC-II Project, Page (1969, p. 1) noted that the economic feasibility of computer grading of essays, made plausible by advancing technology in natural language processing by computer,

would "alter the nature of certain national testing programs, and may eventually change some of the daily conduct of education in most schools and colleges". Page further argued that through this process, "much may be learned about the nature of essay evaluation, the relation of lexicons to subject matter knowledge, and other aspects of psychology and linguistics which are educationally relevant".

Finally, the most important potential for this line of research has been aptly described by Marcotte (1969, p. 8):

> The use of a computer to rate student essay responses will be beneficial for both teachers and students. A fast, effective, and inexpensive method for analyzing and grading essays for both style and content will relieve the teacher of the burdensome and time-consuming task of correcting essays. The student will also profit because he will receive almost immediate feedback on the essays he has written, a not too common procedure in the schools today. Moreover, students will benefit from the extra preparatory time allowed to teachers, which has been seriously limited because of the task of evaluating essays, and from the more effective study demanded by the essay examination.

Summary

This chapter has presented an outline of the main directions the current investigation would take. Major hypotheses associated with certain primary features of the data were described; the rationale underlying the project and the significance of the study for education were discussed.

CHAPTER II

BACKGROUND LITERATURE

Introduction

The area of essay analysis spans several fields of research in the behavioral science, humanities, and computer technology disciplines. In terms of computer technology, perhaps the broadest category under which the analysis of essays is subsumed is that of "natural language processing". This general field includes a number of overlapping areas of inquiry, ranging from the problems of devising the simplest automated parsing algorithms and content analytic techniques to the most complex and sophisticated levels of artificial intelligence programming -- those involving machine interpretation, "understanding," decisionmaking, and response. (An excellent example of avant-garde achievement in the artificial intelligence area is the program developed by Winograd at MIT, described in his recently published Understanding Natural Language.)

In the behavioral sciences, the field of psycholinguistics (incorporating both content and stylistic analysis) is relevant to the present topic, as are psychometric considerations and those other areas of inquiry in psychology, sociology, etc., in which the written artefact is examined in order to draw conclusions about

individual abilities and/or group traits.

Measurement of writing ability is the broad area in the humanities disciplines which most logically contains the boundaries of the present research. Associated with this field of investigation are studies researching the problems of essay test reliability and validity (see Coffman, 1971, for an excellent overview of current work in this area), and explorations of "readability" questions. Computational linguistics and content analytic studies of historical and pure literature documents are also highly related to the work under discussion here. An especially interesting example of the last-mentioned research area is the 1968 content analytic study by Zicky at the University of Connecticut, of passages in Paradise Lost containing speeches by, or descriptions of, the major characters in the poem.

The Page and Paulus Study

The Page and Paulus study, completed in 1968, investigated the ability of 30 stylistic variables to predict student essay grades. Their sample consisted of 272 Wisconsin high school students (grades 8-12), who had each written essays on two different, very general topics. The average essay grade given each essay by eight expert judges served as the criterion measure. The multiple correlation coefficients obtained, for both essays, were about

.71 (.65 after cross-validation). In a subsequent stage of their study, Page and Paulus employed the same predictors in multiple regression equations using five different criterion measures (eight-judge average used to determine each essay score on each trait): (1) ideas or content, (2) organization, (3) style, (4) mechanics, and (5) creativity. A criterion measure representing the average score for these traits was also used. The highest MULTR achieved here was for the "style" criterion (.73); the lowest was attained for the "organization" criterion (.61). Of the 30 independent variables employed, the following predictors were most prominent in the analyses: (1) length of essay, (2) proportion of commas used, (3) proportion of dashes used, (4) standard deviation of word length. In addition, the proportion of spelling errors and the number of paragraphs were significant predictors in two of the analyses.

Further, this lengthy study -- the first in the series containing most of the other studies to be described here -- outlined a number of theoretical and methodological considerations and problems in the area of essay analysis, and provided valuable suggestions for the future direction of research on this topic.

McManus Study

In 1968, McManus completed a content analytic study of college-level history examination essays. His data consisted of four

sets of responses to short-answer essay items written as part of a final examination in a History of Western Civilization course at the University of Connecticut. Eight graduate students in the field of history were employed to rate each essay using a five-point scale.

Three types of independent variables were chosen for use as predictors of the criterion, "essay grade" (defined as the average of the eight scores assigned by the raters to each student response). "Type A" predictors consisted of single isolate words representing key content words in the student responses. These words were chosen in an empirical manner by employing techniques analogous to item selection methods using an item discrimination index. By rank-ordering the student texts on the basis of grade and employing a computer program which identified words appearing in the texts which distinguished the upper 27% of each group of texts from the lower 27%, a set of good "discriminator" words was produced. Each of these became a predictor of essay grade.

The second independent variable category (termed "Type B") contained a single predictor, McManus' so-called "relevancy variable" -- a product formed from the sum of the matches on the chosen "Type A" words and the length of the essay.

Finally, a "Type C" variable category was defined, consisting

of two stylistic variables (number of commas and standard deviation of word length) which had been identified as effective predictors in the Page and Paulus study.

Fisher's SCORTXT program (1968) was used to obtain values on the various predictors in each essay. Eight multiple regression analyses were then performed: one with Type A and Type B predictors and one with Types A, B, and C predictors for each of the four essay items represented in the data. The resulting MULTR's ranged from .57 (Item A, using Types A and B predictors) to .83 (Item D, using Types A, B, and C predictors). In all analyses, the relevancy variable was the best single predictor of essay grade. McManus' results indicated that the stylistic predictors (Type C) added little to the MULTR value.

Marcotte Study

Marcotte's 1969 study, A Computerized Contingency Analysis of Content Graded Essays, made use of the same pool of data that was employed by McManus in his study, but took a very different approach to the essay analysis problem. Computer-usable scoring keys for the twelve essay items represented in the data were developed by taking the teacher keys provided by the essay test author and reducing each to a series of "discrete simple thought units with all function words eliminated" (p. 2). The resulting word strings

representing the reduced text of each key were then arranged in what Marcotte called a "contingency matrix". Each filled cell of the matrix (representing a key word) was really a category containing appropriate synonyms for the content words in the teacher key. A special computer-usable thesaurus of such words was generated for purposes of the study.

Rules were devised for handling certain problems arising from the nature of the resulting "thought units". For example, the following sentence in the teacher, "Dante, Rabelais, and Boccaccio wrote in the vernacular," would yield three simple thought units: (1) Dante wrote vernacular; (2) Rabelais wrote vernacular; (3) Boccaccio wrote vernacular. Since in this case it is the proper nouns that are of interest, and to eliminate the possibility of a student receiving an inflated score because his response had included only the words "wrote" and "vernacular," these sentences were reduced to single words (the proper nouns). These single words, however, were still called contingencies to avoid ambiguity.

A modified version of Fisher's SCORTXT program was employed for the data analysis. Student responses were compared with the stored contingency matrix for an item and cells of a "student response matrix" corresponding to the cells of the key word contingency matrix were filled according to matches made between the

student essay and the scoring key. Two types of scores were generated: (1) total credit, if all categories were matched -- a score of "1"; (2) partial credit, if half or more, but not all, of the categories were matched -- a score of less than one, calculated by dividing the number of correct matches by the number of categories within the contingency. In addition, two types of "student response units" were considered: (1) one-sentence, called "strict" scoring, and (2) three-sentence, termed "loose" scoring. To determine a student score using the loose scoring response unit, each overlapping cluster of three sentences in the student response were examined sequentially to determine matches with the categories in the contingency matrix. Marcotte concluded that, using his procedure, the computer performed as well as the human judges in assigning essay ratings and that the three-sentence student text processing was a better means of predicting human judge scores than was one-sentence processing. While the present study contains no exact replications of Marcotte's procedure, a number of similarities (in terms of handling the content aspects of the data) will be observed between the two approaches.

Content Analyses Using "Vagueness" Concept

Two 1969 studies -- Miller, Marcotte, and Martin; and Miller, Fisher, and Kaess -- explored the relationship between an operat-

tionally defined concept of "vagueness" and student achievement scores. In the former study, the writing traits "opinionation," "vagueness," and "specificity" were used as predictors of essay quality. After shrinkage and attenuation, a MULTR of .47 was obtained using these predictors.

In the latter study, the investigators attempted to predict high school student achievement (on a multiple-choice examination) using measures obtained from an analysis of their teachers' lectures. The vagueness factor identified in the lectures correlated negatively with the criterion in both sets of lectures studied. A verbal fluency factor displayed lower, but significant criterion correlation and consistency. Factors indicating amount of information and interest achieved statistically significant correlation with the criterion in the second lecture but not in the first.

Whalen Study

A study by Whalen (1970), utilizing a sample of seventh grade student essays, sought answers to the following questions: (1) Do the linguistic variables used successfully by Page and Paulus, and Janzen (1968) serve equally well as predictors of writing ability for average seventh grade students? (2) Can the prediction of mechanical accuracy be improved through the use of additional variables and by achieving a more objective and relia-

ble criterion measure of mechanics? (3) Can computer evaluation of essays predict with any degree of success student scores on a standardized test of English mechanics, including capitalization, punctuation, word usage, and spelling?

A total of 26 predictor variables (both stylistic and mechanics-related) were used in the analyses, including 17 stylistic variables from the Page and Paulus study. Three major dependent variables were employed: (1) overall essay grade; (2) mechanical proficiency -- defined negatively in terms of the total number of mechanical errors in the student essay; (3) language ability -- defined as student scores on the "Mechanics of English" and "Spelling" subtests of the California Language Test. In addition, scores on the following four subtests of the CLT served as criterion measures in certain of the multiple regression analyses which were performed.

Although the original MULTR values (for the "full" models -- those in which all 26 predictor variables were employed -- these values ranged from .79 to .91) shrank drastically in the process of cross-validation, Whalen was able to draw conclusions from his results which led to affirmative responses to the initial questions posed in his investigation. He found that, in general, the predictor variables used in the Page and Paulus study did appear to

contribute substantially to the prediction of writing ability at the seventh grade level. In addition, several of the "new" predictors identified by Whalen (including number of capitalization errors, number of occurrences of the word "then," and number of usage errors) were important contributors to the prediction of mechanical proficiency. Whalen also concluded from his results that the "mechanics model" was the most reliable of the three models tested, and that machine prediction of total test scores for the California Language Test from a single essay was just as successful as the prediction of that same essay's overall grade.

Archambault and Greene Studies

Archambault (1969) and Greene (1970) each developed algorithms for predicting scores on the Torrance Test of Creative Thinking. Archambault's investigation concentrated on the first three "activities" in the test, while in Greene's analysis, responses to the last four test activities were the data. The responses on each activity by a sample of students in grades 4 through 7 were scored by both trained student and "professional" raters on three dimensions of creativity identified by Torrance: fluency, flexibility, and originality. These studies were essentially content analytic, requiring the construction of elaborate "dictionaries" to repre-

sent categories provided by Torrance as the basis for scoring on the three dimensions.

The results of these two studies were quite successful. In Archambault's analysis, only two of the nine MULTR's calculated (one for each of the three dimensions on each of the three activities) fell below .90. Greene obtained MULTR's for the fluency dimension ranging from .92 to .99. The range in his data for the flexibility dimension was from .84 to .91, and for originality, from .73 to .87. Using restricted models (employing no more than half of the original set of predictors), Greene obtained results which paralleled those using the full models -- i.e., there did not appear to be a significant loss in predictive power when the small subset of effective predictors was employed in place of the full set of original predictors.

Summary

The preceding discussion has considered the relevance of the topic of essay analysis to a wide variety of disciplines. Studies describing research work over the past five years on various aspects of essay quality were briefly outlined, and their relationships with the current investigation were emphasized.

CHAPTER III

PROCEDURES

Introduction

The first section of this chapter describes the nature and rationale of the sample used, and the procedures used for the collection of data for the AEC-II Project. The second section of the chapter deals with manual and computer strategies employed in the content analysis of the sample essays. The final section of the chapter presents a restatement of the hypotheses underlying the study, together with the methods used for testing these hypotheses.

Data Collection Procedures

Rationale

The design of the AEC-II Project called for a sample of 48 essay sets -- twelve sets for each of four subject areas. Twenty-four classes would be represented in the sample, each class submitting essays on two different essay questions. The range of grade levels in the sample would be from seventh through twelfth grade.

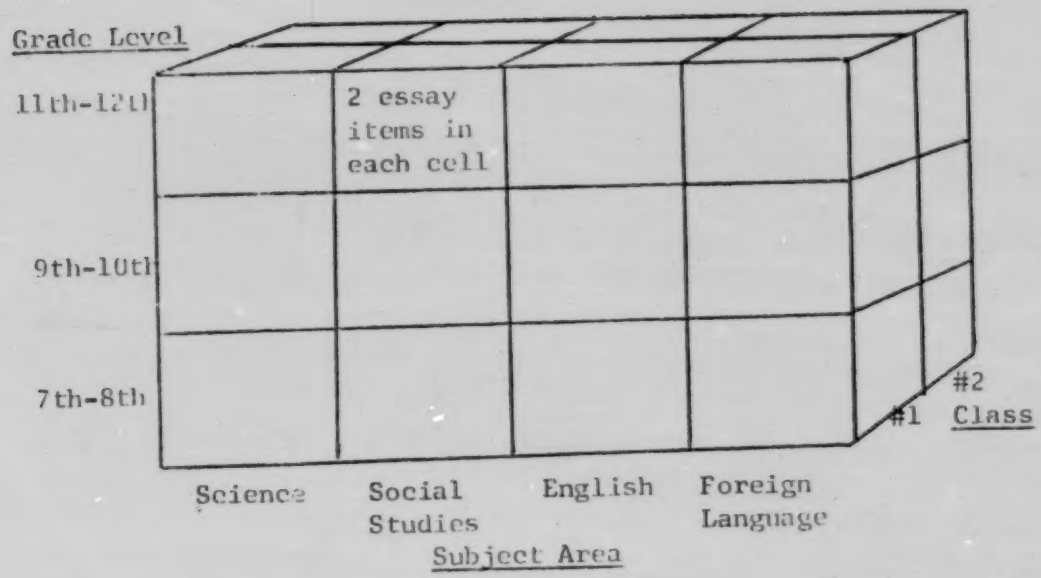


Figure 3.1
Data Collection Design for AEC-II Project

The essay questions and student responses of which the sample data is composed represent typical classroom essay examinations given at the junior high school and high school levels. During the 1969 school year, Mr. Russell Agne of the Bureau of Educational Research (BERS) at the University of Connecticut began the data collection procedure for the Project by drawing a random sample of teachers from a list of all Connecticut public secondary school teachers of science, social studies, literature, and foreign language. This listing, containing 7,587 names, was obtained through the coopera-

tion of the Connecticut State Department of Education. A random numbers table was used to select 200 names from this list. Teachers whose names had been selected were called during late May, 1969, in the order in which their names had been drawn from the list using the random selection procedure.

Each teacher called was informed that student responses to typical classroom essay examinations were being solicited from them for a federally-funded research project "which is attempting to evaluate student responses to essay questions" (Agne, 1969, p. 11). If the teacher volunteered to contribute a set of essays, and planned to submit essays written on examinations to be given at a future date, he was requested to be sure that the test was given under ordinary conditions, with no special notice given to the students that they were participating in a research project. No essays written in response to "take-home" or "open-book" examinations were used. Each teacher was requested to provide, with each essay set, (1) a copy of the question generating the essays, (2) the grade given to each essay, (3) a teacher key for each question, and (4) a "model student essay" -- that response which the teacher felt to have been the best one contributed in the class. Of the teachers called at this time, 32 volunteered to have one of their classes contribute a pair of essay sets for the Project.

Only one class per teacher was used.

During each telephone conversation, the teacher was asked to describe the classes he taught in which essay examinations were given as a part of the usual instructional procedure. The interviewer then rolled a die to select randomly one of these classes for inclusion in the Project.

In the fall of 1970, the sixteen additional essay sets needed to fill the proposed data matrix were obtained by Mr. William McCook of BERS, using the procedure described above. In all, the data consisted of 831 essays, the average essay length being 60-70 words.

Essay Handling Procedures

Following agreement by a teacher to participate in the Project, a prepaid mailing envelope, addressed to BERS, and a covering letter was sent to each volunteering teacher. A modest honorarium of $15.00 was paid to each teacher for his help. As each set of essays was received, it was xeroxed and the originals returned to the teacher. Since some essay examinations submitted contained as many as ten separate essay questions, a random selection procedure was employed in these cases to identify two questions per class from the examination for subsequent analysis. Not all teachers submitted keys with their questions and in some cases no model

essay had been chosen. In no instance were both the teacher key and the model student essay missing for a question.

For purposes of keypunching the essays and the teacher keys, the instructions provided by Fisher (1969) in The SCORTXT Program for the Analysis of Natural Language by Computer were used. These instructions call for a "text card deck" consisting of: (1) a text header card, (2) any number of text data cards, and (3) an end-of-text card. The text header card contains a text identification number in columns 1-5; this number includes a subject area code, a class code, a "question" code (#1 or #2 in the pair for a class), and a student number. The remainder of this card is ignored in processing and may contain any information desired. The text of the student essay is punched on the text data cards, in columns 1 through 72, with columns 73-80 being used for identification and consecutive numbering of the cards. In punching the data, card boundaries are ignored; words not fitting wholly onto one card are continued starting in column 1 of the following card. Words are separated by one blank, sentences by two consecutive blanks, and paragraphs by four. Italics are indicated by the use of the symbol, "__", preceding each word to be italicized. Consistency in following these instructions is vital in subsequent computer processing to insure the success of computer algorithms developed to

recognize such features of the text as word, sentence, and paragraph boundaries.

The coding system used in keypunching the AEC essays is shown below:

TABLE 3.1
CODING SYSTEM FOR IDENTIFICATION OF AEC ESSAYS

Item	Column	Digits	Meaning
Grade Level	73	1	Grades 7 or 8
		2	Grades 9 or 10
		3	Grades 11 or 12
Subject Area	74	1	Science
		2	Social Studies
		3	English
		4	Foreign Language
Class	75	1,2,3	Sequential order in which a set of essays within a cell was keypunched
Student Number	76-77	1,99	Identification number of a given student within a given class
Question Number	78	1,2	Identification of question in a class pair
Card Number	79-80	1,99	Card identification for a given student answer

After keypunching, two printouts were made of each set of data cards, and the keypunchers proofread their work against the students' papers.

Essay Rating Procedures

Chapter V of the Analysis of Essays by Computer (Page and Paulus, 1968) provided the model which was used for the conduct of the essay rating sessions. Sixteen secondary school teachers (four in each subject area), teaching within a 50-mile radius of the University of Connecticut, were selected using a modified random selection procedure for participation in the essay rating sessions. Secondary schools in the vicinity were called according to a random schedule and teachers recommended by their department chairmen as having particular competence in the teaching of their subject were asked to volunteer to participate. Only one teacher was solicited from each school contacted. Each teacher was paid an honorarium for his participation, the amount varying on the basis of whether the rating took a half-day or a full day. In most cases, the length of time necessary to complete the ratings on an entire set of student essays for a subject area (six classes' responses to twelve essay questions), was a half-day per rater for the science and foreign language essays, and a full day per rater for the English and social studies essays.

The rating sessions were carried out on two separate days (one in October 1971 and the other in March 1972). The specific instructions given to the judges are shown in Appendix A. Eight teachers were present at each session. Interaction among the teachers during the sessions was not encouraged and very little occcurred. The computer program that was used to print out copies of the student essays in a format suitable for rating purposes randomly ordered the presentation of the twelve essay questions to each rater; the program also randomized the order in which the student responses to any one question were presented to a judge. Judges were asked to rate each essay on a 1-to-5 scale ("1" being the highest grade) and to use the entire range of scores in evaluating the essays for a given question. The judges were also requested to evaluate the teacher key and the model student essay provided, both before attempting to rate the student essays for a question and after rating them. The following scale was used for this purpose: (1) quite helpful and complete, (2) of considerable use, (3) of some use, and (4) of almost no use at all.

Computer Strategies for Content Analysis

Introduction

Three broad techniques seemed required for developing compu-

ter-usable keys for matching with the student essays in the data:

1) the reduction of teacher keys and model student essays to sets of short elementary propositions, chiefly by breaking up the sentences of the keys and model essays into phrases representing a single thought and eliminating the function words in the phrases;

2) the recognition of syntactical transformations of these phrases in the student essays;

3) the recognition, in the student essays, of synonyms of the words appearing in the key phrases.

As a first approximation to an efficient strategy for achieving (2) above, all words in both the student essays and the computer-usable keys were reduced to a root-word form, chiefly for the purpose of making a word in the student essay recognizable as a match for a word in the computer key, regardless of the part-of-speech inflection which the student may have chosen for the word. Also, this strategy allowed for any permutation in the word order of a key phrase when searching for it in the student essay.

The figure below shows a diagram of the general strategy employed in the content analytic phase of the study.

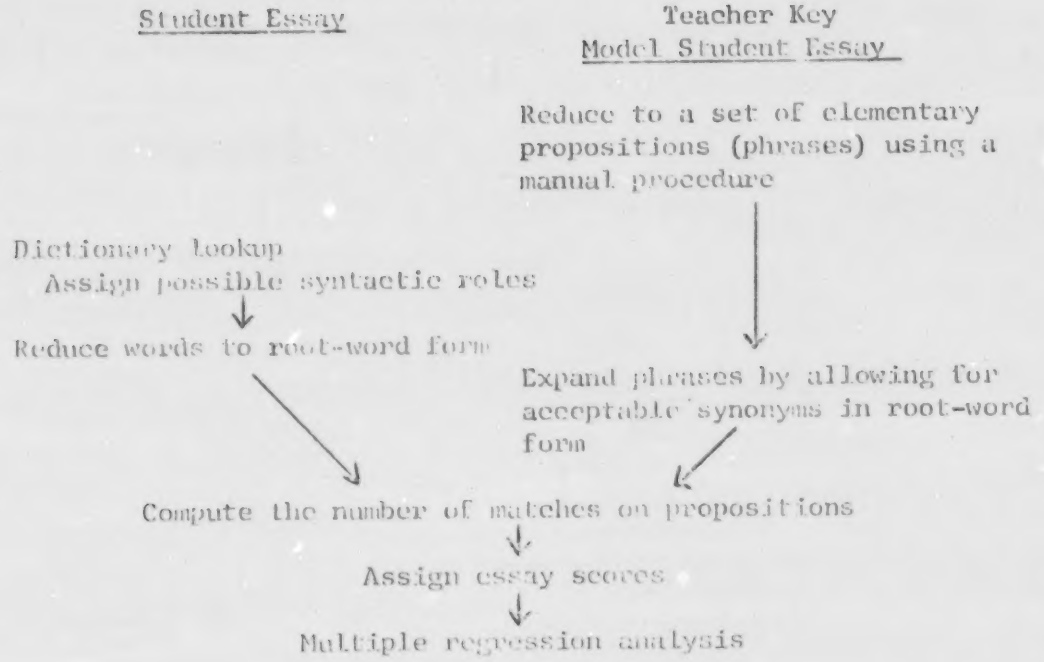

Student Essay
Teacher Key
Model Student Essay
Reduce to a set of elementary propositions (phrases) using a manual procedure
Dictionary lookup
Assign possible syntactic roles
Reduce words to root-word form
Expand phrases by allowing for acceptable synonyms in root-word form
Compute the number of matches on propositions
Assign essay scores
Multiple regression analysis

Figure 3.2
General Strategy for Content Analysis

The computer analysis of student essays to be undertaken in the AEC-II Project was originally envisioned in terms of a strategy consisting of four stages, each stage requiring its own group of computer programs. The input to each stage would be provided by the output of the preceding stages. The tasks assigned to each stage were as follows:

Stage 1. From the teacher key and/or model student essay for each question, extract a list of key phrases in root-word form.

Stage 2. Expand the list of key phrases to incorporate accep-

table synonyms for the words in the key phrases.

Stage 3. Reduce the student essays to a simplified form by converting the words to root-word forms and by deleting most function words and punctuation marks.

Stage 4. Match the simplified student essays against the expanded list of key phrases for each question, outputting counts of the number of matches on these phrases for each essay.

The algorithms and programs for these stages were actually developed in reverse order, the stages as listed being in descending order of difficulty. In the present study, certain of the tasks necessary for Stage 2 and all of the tasks related to Stage 1 were done manually. Note that the strategy does not employ any parsing in the conventional sense. The simplest structure for each stage was implemented for the purposes of the present study.

All computer programs for completing Stages 2 through 4 were devised and documented by Mr. Peter Tillett, a Graduate Assistant in the Bureau of Educational Research.

Stage 1

This stage consisted of extracting, from the teacher key (where available) for a question and the model student essay (where available) for that question, a list of key phrases in root-word form. All the necessary work for this stage was accomplished man-

ually, without the direct aid of computer programs, by the author, without any prior knowledge of the actual student responses which had been written for the question. Computer algorithms were not used for this stage because the basic problem here -- the recognition of phrases as meaningful groups of words, and the selection from them of key phrases -- is very difficult to accomplish by computer. The former operation would require a minimally powerful parsing algorithm, the latter a semantic capability in the programs.

The manual procedures used were straightforward and simple in concept, though often difficult in application. The sentences of the teacher key and the model student essay were examined and meaningful phrases within them were recopied, eliminating all function words. The phrases thus generated were then reexamined and revised, if necessary, for the purpose of conveying the thought of the phrase as simply and unambiguously as possible. Redundant phrases were eliminated. In a very few cases, additional phrases, taken from sources other than the teacher key and the model student essay, were prepared by the author, if the thought conveyed by such a phrase seemed particularly relevant to the question being asked, though not directly mentioned in either of the two main sources for the phrases. Such phrases were kept to an absolute minimum, however, since the study was mainly concerned with what could be

accomplished with the data already provided, however inadequate the data might prove in certain cases.

Lists of computer-usable key phrases were prepared for 44 out of the 48 essay sets. The essays associated with four items in the English area had to be deleted from the study because the responses called for by these questions appeared so general that no single list of meaningful phrases could be generated from each such question as it had been presented.

As an example of the type of question for which no key phrases could be framed, one of the English essay questions deleted from the content analysis is presented below:

> Question: Please answer the question in clear argumentative writing. (State your position and justify it.) State what you believe to be the basic problem in your area of study. Justify your statement with at least four strong examples from your research. Be sure to explain what each example shows about your statement ("I believe the basic problem with _____ in our society is _____.").
>
> Teacher Key: This test is based upon individual research papers written by the students. There is no key as such. Each student's paper is judged according to his research topic and the degree to which he discusses it on the test. The common requirement is that each question be answered in argumentative writing with a specific number of examples to justify the position of the student's response.

The total number of essays ultimately available and acceptable for processing was 758.

As work progressed on this stage and on Stage 2, the desira-

bility of limiting the number of phrases used for a question became more and more apparent. Beyond a certain point, one found that the additional phrases being generated introduced more ambiguity than relevant information. Each phrase was intended to carry a single thought; therefore, once function words and ambiguous or redundant words had been eliminated from a phrase, the number of words in it were few, ranging from one to ten. The number of phrases generated for a question ranged from one to thirteen, the average number of phrases per question being five.

Stage 2

The need for Stage 2 in the computer analysis resulted from the recognition that students might employ in their essays acceptable synonyms for words appearing in the computer-usable key phrases, and from the recognition that the order of words in a key phrase could be acceptably varied.

Manual Procedures: For the purposes of the present study, a major portion of the tasks related to Stage 2 was done manually, chiefly because of the lack of a manageable-sized, computer-usable thesaurus. The procedures were as follows:

1) Once the computer-usable key phrases for each essay in a particular subject area had been devised, each word appearing in a key phrase was located in Roget's International Thesaurus, 3rd

edition and a list of synonyms for each word was prepared. Each such synonym list was selective in that only those words likely to be part of the vocabularies of U.S. students at the grade levels covered by the study were included. This was done to shorten subsequent computer processing time; synonym lists for many of the words appearing in the keys would otherwise have been quite extensive. The following table shows the thesauric categories listed in Roget for the word "land" (a word which appears in the key phrases); had the total number of single-word synonyms for "land" been included in the special thesaurus being prepared for the key words in the phrases, there would have been a synonym list of 389 entries for this word alone.

TABLE 3.2
THESAURIC CATEGORIES AND NUMBER OF SYNONYMS
APPEARING IN ROGET'S THESAURUS FOR THE WORD "LAND"

Category Label	Entry Code	Number of Synonyms
Nouns		
Region	179.1	27
Country	180.1	46
Ground	384	147
Real Estate	808.7	21
Make Land		
Navigation	274.90	5
Land	299.8	16
On Land, By Land	384.11	9

TABLE 3.2 continued

Category Label	Entry Code	Number of Synonyms
Verbs		
Land a Plane	277.49	27
Disembark	299.8	16
Alight	314.7	11
Wangle Into	646.23	24
Catch	820.14	30
Land On	1008.12	10
		389

If a synonym found in Roget for a word was an idiomatic phrase, or a slang expression, it was not, as a rule, included in the synonym list. Each word in a thesauric class was considered for inclusion in the synonym list, regardless of its part-of-speech.

2) After synonym lists had been prepared for each word appearing in the computer-usable keys, each synonym was checked against the dictionary of words used in the student essays (one output of Stage 4). If the synonym could not be located there, it was deleted from the list. Because of this deletion process, the number of synonyms ultimately appearing in the AEC thesaurus for the word "land" (the example used above) was only eight. The dictionary of words used in the student essays was also used to reduce all words in the synonym list to root-word form. Both of these steps were taken for the purpose of making the subsequent computer runs less time-consuming and more efficient.

3) Finally, each word in each synonym list was punched, one word per card, according to the format required for the computer run for Stage 2. Each computer-key phrase word served as the category heading for its list of synonyms and was designated as such by punching a "1" in column 80 of the first card on which the word appeared. Because of the structure of the computer program for this stage, it was necessary to prepare two cards for each computer-key phrase word, one to designate the word as a category heading and one to enable the computer to recognize the word as a synonym of itself when attempting to match words of the student essays with the computer-key phrases. These category heading cards, each followed by the cards containing their respective synonyms, formed the "thesaurus" used as input to the computer program for Stage 2.

Stage 2 Computer Program Description: Input to Stage 2 consists of teacher keys and model student essays reduced to sets of short elementary propositions or phrases. Output from Stage 2 is these phrases: (1) expanded to allow for synonyms of the words in the phrases; (2) modified so that all words in the output are in "root-word form"; (3) structured in the "ring format" required as input to Stage 4. A supplementary input to Stage 2, required to achieve Step 1 above, is a thesauric dictionary, giving synonyms for the words employed in the computer-key phrases. The program

for this stage consists of approximately 350 FORTRAN statements.

Stage 3, Run 1

The computer runs necessary to achieve Stage 3 are described in Figure 3.3. In the first run, the program REFORMAT takes as input the student essays and a list of function words and reduces the essays to a simplified form by breaking up the essay text into words and punctuation marks and outputting these as separate fixed-length records with a sequencing variable assigned (for later re-sorting in Run 4). The list of function words included as input to Stage 3, Run 1, is presented in Appendix B. This list does not by any means exhaust all possible entries under each category of function words described below, since it was prepared using only words actually appearing in the student essays. For the purpose of creating this list, a "function word" was defined as any article, personal pronoun, conjunction, relative pronoun, interrogative, or preposition (providing the word could be identified as a preposition without reference to context). Other input material (questions, teacher keys, model student essays) is passed over by the program. The purpose of Run 1 is to enable the words employed in the essays to be sorted into alphabetical order (Run 2) to facilitate the matching of words against a magnetic tape dictionary (Run 3).

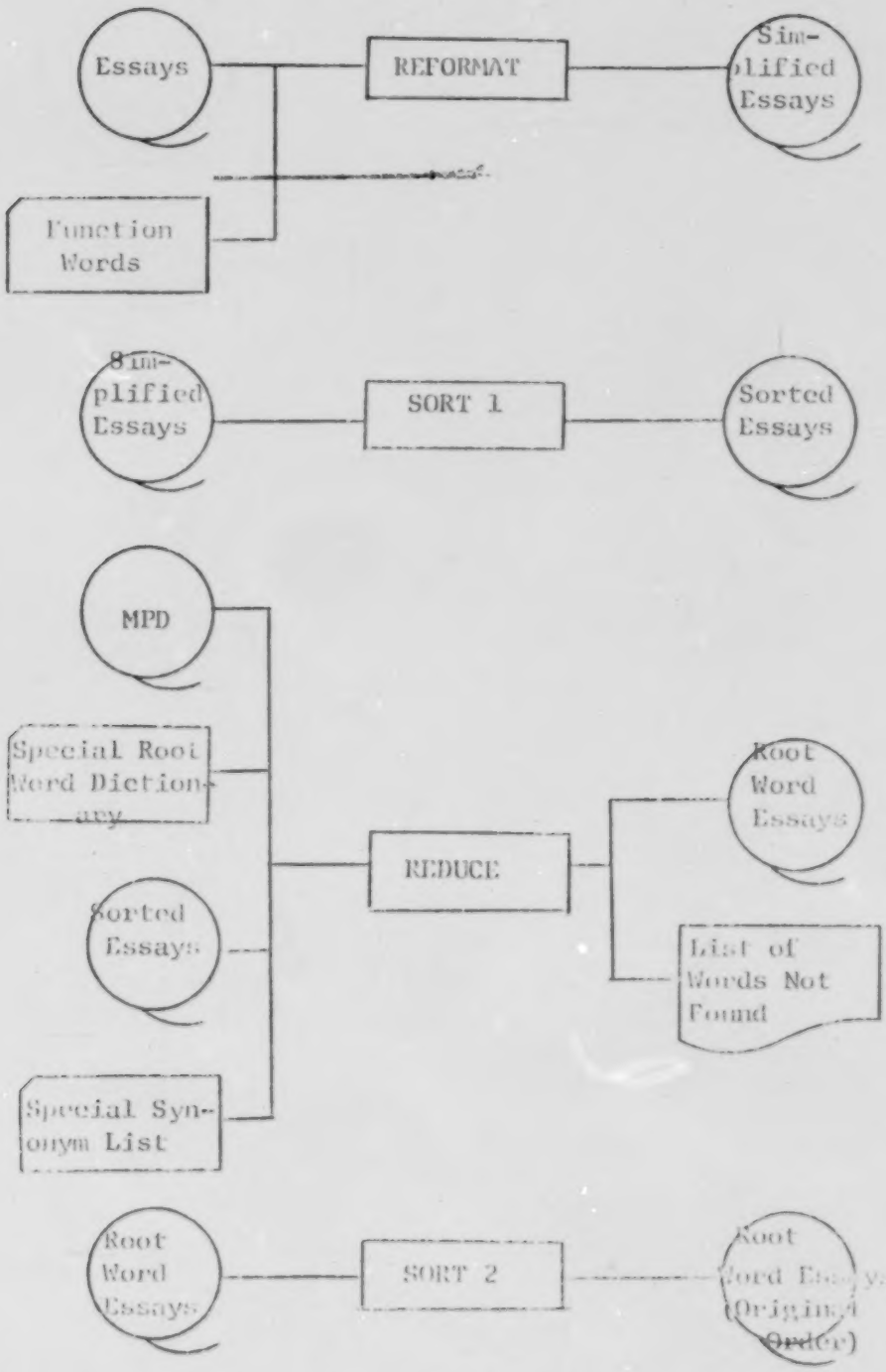


Figure 3.3
Runs Involved in Stage 3

The program consists of three subroutines. The MAIN subprogram bypasses the question, key, and model for each essay set and reads the essays one at a time into array SA. At the end of each essay, subroutine SAOUT is called to process it. SAOUT breaks the essay up into individual words. A "word" is regarded as a string of non-blank characters preceded by a blank or a punctuation mark. Every punctuation mark is regarded as an individual record, or "word," with two exceptions:

(1) The 0-8-2 character is ignored whenever it occurs. (This character was used to expand the student essay texts as necessary in order to make corrections in the text without having to repunch all the following cards of the text.)

(2) If the characters "," or " ' " occur at the beginning of, or within a word, they are not separated from it. (Retention of the former character allows numbers such as "5,000"; retention of the latter character allows words such as "it's" to remain intact.)

At the end of each word, subroutine WOUT is called to write it out, blank the output holding area, increment the record count, and reset various indicators.

The dash was frequently used in the essays of the sample to set off an appositional phrase -- as in this example. In most

cases, the dash was punched without a blank preceding or following it, thus making it indistinguishable to the program from a hyphen. Since inspection of the essays suggested the use of hyphenation to have been very limited, it was decided to have the program assume every occurrence of the dash as appositional punctuation. Because of this, any hyphenated words included in the computer-key phrases or in the sources for them had to be considered as "idioms," i.e., punched as separate words (no hyphenation) on a single card. Such groups of words would then be treated as single words during phrase analysis.

Stage 3, Run 2

This program used the IBM Library Program SORTD to sort the output of Run 1 into "alphabetical order"; ascending order is specified. This operation is carried out to facilitate the dictionary lookup undertaken in Run 3.

Stage 3, Supplementary Run 2A

This program accepts as input a magnetic tape (from Run 2) which contains the words of all the student essays, alphabetically ordered, and prints out each unique word once, with its frequency of occurrence. The program also yields a count of the total number of essays, the total number of words used, and the total number of different words used. The program consists of 33 statements

with no subroutines. It reads input, accumulating a frequency count and a total word count until a new word is reached. Data for the previous word is printed out, the new word saved, the count of different words incremented, and the frequency count reset to 1.

Stage 3, Run 3

The third program related to Stage 3 is concerned with the reduction of words to root-word form, and the assignment of parts-of-speech to these words (the part-of-speech information is not, however, made use of in the present study). The following input is involved:

Dictionaries

1) Output of Run 1 (an alphabetical listing of words appearing in the student essays)
2) A special subset of words from the Merriam-Webster Pocket Dictionary (MPD) on tape
3) A proper noun dictionary (approximately 200 words)
4) A spelling error dictionary (approximately 800 words -- input here consists simply of the incorrect and correct spelling of each word which was misspelled in the student essays)

Stemming Algorithm

The "Standard Dictionary": The standard dictionary for the AEC-II Project, incorporating only those spellings appearing in the student essays, was derived from the MPD, using both computer programming and manual editing. For each spelling in the MPD, a root-word form of the word and a part-of-speech (with up to three alternative parts-of-speech) are given. This dictionary was employed in the analysis, not only because it provided syntactic information on the words, but because it enabled the use of a greatly simplified stemming algorithm (see below). The algorithm achieves virtually complete success because of the powerful requirement that a "suffix" be removed only if the resulting word occurs in the dictionary with a compatible part-of-speech.

The Proper Noun Dictionary: This list, prepared by hand, contains the approximately 200 proper nouns appearing in the AEC-II essays. This dictionary was included in order to provide the correct part-of-speech code for these words, since they would not be likely to be included in the MPD; this dictionary also enables a count on the proper nouns appearing in each essay, and such a variable was included as a predictor in the multiple regression analyses.

The Spelling Error Dictionary: This dictionary, prepared by hand, includes all the obvious misspellings contained in the AEC-II

essays. Input here consists simply of the incorrect spelling followed by the correct spelling. Since this dictionary was prepared by examining the output of Run 1 (the alphabetical listing of the words contained in the essays), and in consequence the words were observed out-of-context, spelling errors in which the word is a misspelling in the context used, but could be a correct spelling of a word in another context, could not be detected. (Examples: "there" for "their"; "magma" for "Magna".)

Choice of Root-Words: The choice of a root-word spelling for a word was not of great consequence, since the root-word spelling is used merely as a unique class identifier and could equally well be a word or a number (as is used in the General Inquirer program). The root-words, however, were chosen fairly systematically. In the majority of cases, the root-word chosen was the shortest spelling of which the given spelling might be regarded as an inflectional form. This root-word was normally achieved by removing the suffix(es) contained in a given spelling. Prefixes were not removed in order to form the root-word. The shortest spelling was not chosen to be the root-word in the few cases in which the use of the shorter word was very rare compared to the use of the longer spelling, e.g., "opine" vs. "opinion". The infinitive form was chosen as the root-word for verbs in all but one case ("leaves" -- for which the given spelling incorporated more meanings, i.e.,

could equally well stand for the plural of "leaf" or the present singular tense of "leave"). All function words were given a "blank" root-word.

The Stemming Algorithm: The stemming algorithm was applied to all words not located in the MPD. Successful stemming yielded a word appearing in the MPD from which the root-word form could be taken. Syntactic information was specified by the dictionary; however, determination of part-of-speech was partly algorithmic. If, for example, removal of the suffix "-s" yields a noun in the MPD, the syntax of the word is "noun plural"; if it yields a verb, syntax is "verb, third person singular, present tense". If removal of the suffix "-er" yields an adjective, syntax is "comparative adjective"; if removal yields a verb, syntax is "noun". Appendix C shows the suffixes handled by the stemming algorithm, with examples.

The program for this run consists of a main program and fourteen subroutines. The program first reads the proper noun set and sorts it alphabetically, internally. The dictionary is then read and merged with the proper noun set. The spelling error set is then read and sorted alphabetically. Then each word from the student essays is processed. First, the word is looked up in the spelling error set. If found there, it is replaced by the correct

spelling identified in the set, for subsequent processing. The word, corrected if necessary, is looked up in the dictionary file. Regardless of the result, the stemming algorithm is also applied, and the output root-word and syntax information is a combination of the dictionary and stemming algorithm outputs, if any.

The program produces a listing of those words for which the lookup-stemming procedure is unsuccessful; every occurrence of such a word is printed out, together with the essay set, essay number, and word position at which it occurred. The proper noun set, the standard dictionary, and/or the spelling error dictionary may also be printed as output. For each word in the student essays, the program provides the following information: (1) the word as it appeared originally; (2) the correct spelling of the word, if it was misspelled; (3) the "stemmed" spelling of the word (if processing of the corrected word was successful, and if at least one syntax for the word required the use of the stemming algorithm); (4) the root-word for the stemmed word, as supplied by the standard dictionary; (5) part-of-speech codes (up to four different alternatives) for the word, obtained from the standard dictionary.

Stage 4

This stage employs one program, which accepts as input sets of essays (Stage 3 output) and sets of computer-usable key phrases

(Stage 2 output) on magnetic tape. The program determines whether any of the key phrases occur within the search unit designated (in the case of the present study, this unit is either the essay as a whole, or a single sentence within the essay), for essays and key phrases in the same set. Additional input consists of a control card setting the value for six options and brief titles or questions for the essay sets.

The output is controlled by the options, and may be any or all of the following:

The title or question for each essay set

The list of computer-usable key phrases for the essay set

The essay text for each essay

The key phrases matched within each essay

The "matching scores" assigned to each essay

Punched cards containing the scores for each essay

A phrase is any sequence of specified words. A phrase is considered to be present in a unit if the words may be found within the unit in any sequence, regardless of whether other words in the unit intervene between those in the phrase. For sentence units, the boundary-markers are the beginning of the essay, the end of the essay, and the symbols ".", "?", and "!". The program allows for phrases up to 25 words in length, but this limitation may be

easily modified.

A "word" is any set of symbols, including blanks, up to twenty symbols long. These may be English or foreign words, or numbers, or part-of-speech codes, etc. (Thus, the program could as easily search for parts-of-speech codes as it could words, depending on the form in which the essays and key phrases were input.)

The program is also able to recognize "idioms" -- i.e., sets of words to be regarded as single words for matching purposes (for example, "World War II"). For a key phrase idiom to be matched within an essay, the identical sequence of words in the idiom must occur within the search unit without any intervening words. The symbols ".", "?", and "!" may be part of an idiom; any such symbol is first considered as a potential part of an idiom -- if this test fails, it is considered end-of-sentence. Idioms of lengths 2-5 words are presently provided for by the program; this situation, however, may be easily altered. Idioms may occur at any point within a key phrase.

The program merely counts the number of matches within each essay (searching through one specified unit at a time) on words in each key phrase. The program consists of a main program and eight subroutines.

Sample output from Stages 2 through 4 is displayed in Appendix D.

Hypothesis-Testing Procedures

The major hypothesis under consideration in the study was that the inclusion of certain content variables with certain stylistic variables in a multiple regression equation, using the average grade for each student essay as a criterion, would significantly improve the MULTR value obtained when only the stylistic variables were employed as predictors. Based on the results of previous studies (Page and Paulus, 1968; McManus, 1968; Marcotte, 1969), the following eleven stylistic variables were chosen for use in the multiple regression analysis of the student essays:

length of essay (LENGT)[1]

spelling errors (SPELL)

proportion of prepositions in essay (PREP)

proportion of proper nouns in essay (PROPN)

average word length (MEANW)

standard deviation of word length (SDW)

average sentence length (MEANS)

standard deviation of sentence length (SDS)

proportion of commas in essay (COMMA)

---

[1]The labels in parentheses are those used to denote the variables in the tables appearing in Chapters IV and V, and in the appendices.

proportion of dashes (hyphens) in essay (DASH)

proportion of relative pronouns in essay (RELPR)

Six content-related variables were designated for inclusion in the regression analysis. These were:

proportion of teacher key source phrases completely matched in essay (KEYC)

proportion of teacher key source phrases partially matched in essay (KEYP)

proportion of model student essay source phrases completely matched in essay (MODC)

proportion of model student essay source phrases partially matched in essay (MODP)

proportion of "other" source phrases completely matched in essay (OTHC)

proportion of "other" source phrases partially matched in essay (OTHP)

The various sources for each computer-usable key phrase were identified at the time the key phrases were prepared, a "source code " being included with each phrase. The reader will recall that in the vast majority of cases, the computer key phrases were based directly on the teacher key and/or the model student essay; in only a few instances were phrases created in which the idea represented by the phrase did not appear in either of these two sources. Because of the small number of such phrases (the "other" source category), difficulties were encountered in the initial

computer runs of the regression analyses, resulting in the deletion of the "OTHC" variable from the study.

In many instances, a computer key phrase is based on an idea presented in <u>both</u> the teacher key and the model student essay. In these cases, the matches obtained on the phrases could influence the resulting value of more than one variable (e.g., both MODP and KEYP).

All predictor variables were transformed ($\bar{X}$=50, s.d.=10) to standardize the variables prior to the multiple regression analysis.

To test the hypothesis that gains in MULTR values resulting from the addition of the five content variables to the multiple regression equations would be greater for the science and foreign language essays than for the English and social studies essays, the differences between the two MULTR's (the one using stylistic variables only and the one obtained using the full set of sixteen predictors) were compared. The same procedure was followed in testing the hypothesis that gain in MULTR value resulting from the inclusion of the content variables would be greater for the 7th-8th grade level than for the 11th-12th grade level.

Using the full set of sixteen predictors, two MULTR's were calculated for each analysis level -- one using the essay as the

search unit for content variable values and one using the sentence as the search unit. These two values were compared in order to test the hypothesis that the choice of search unit would not significantly affect the MULTR level when all sixteen predictors were used.

Content Specificity: Since the essay questions varied greatly, both across and within subject areas, in terms of the relative emphasis the teacher seemed to be placing on matters of content and style, it was decided to include some measure of the "content specificity" of each question as a basis for grouping the student essays for comparison purposes. In order to obtain measures of content specificity on the essay questions, a page containing (1) the text of the essay question, (2) the teacher key, and (3) information on the grade level of the students responding to the question was prepared for each question. A package containing such information for all essay questions in a given subject area was made up for each of five judges for that subject area, the order of presentation of the questions being randomly varied in each instance.

Graduate students with teaching experience taking courses at the University of Connecticut were asked to rate the essay questions in the sample for content specificity. Some raters

completed ratings for all four subject area packages, while others rated only questions in those subject areas in which they had teaching experience. Appendix E displays the instructions given to these raters and a sample of the rating form used. A ten-point scale was provided for rating each essay question.

It was hypothesized that if the essay questions were rank-ordered on the basis of the average content specificity rating given the question by the five judges, and the student essays sorted into three groups (high, middle, and low) using the relative content specificity of the questions as a basis for sorting, that multiple regression analyses would reveal a greater gain in MULTR value for the high content specificity group than for the low content specificity group when the content variables were included in the multiple regression equations.

Summary

The data for the AEC-II Project consisted of two sets of essays drawn from each of 24 classes representing four different subject areas and six grades (regrouped into three levels for analysis purposes). The data collection method employed and the procedures used for handling the data were described in this chapter. In addition, the rationale underlying the content analytic portion of the study, and the computer algorithms developed for the content analysis were briefly outlined. Finally, the

hypotheses to be tested in the study were restated and the hypothesis-testing procedures used for each were indicated.

CHAPTER IV

RESULTS

This chapter is divided into two major sections. In the first section, a number of descriptive measures calculated for the sample data are presented. The second section begins with a brief description of the calculations used to determine the various values reported in the multiple regression analysis results. Following this introduction, the four analysis series performed in the study are presented and the results of each series reported and summarized. An analysis "series" was defined by the basis on which the essays in the sample were grouped for analysis. The four series were:

- Analyses Using All Essays
- Analyses by Subject Area
- Analyses by Grade Level
- Analyses by Content Specificity Level

The interpretation of the results of these analyses and a discussion of the conclusions to be drawn from them are presented in Chapter V.

Descriptive Measures

Judges' Ratings of the Essays

Means and standard deviations of the judges' ratings of the student essays, by essay question, are shown in Tables 4.1 through 4.4, each table representing a different subject area. The item numbers presented in the tables specify each item in terms of the grade level in which it was used (left-most digit), the subject area represented by the item, the class (within subject area), and the essay question (#1 or #2) within class. The highest average grades (i.e., the lowest mean values, since a score of "1" was the highest grade an essay could receive, while "5" was the lowest grade) were given in the science area, while the lowest average grades were given in English. Although the average intercorrelation between any pair of judges was not high, as shown in Table 4.5, the pooled judge reliability for each subject was quite adequate, ranging from .785 (English raters) to .889 (foreign language raters), as indicated by Table 4.6. To obtain the reliability estimates presented in this table, a computer program was used which calculated reliabilities for the total group of judges in each subject area and for all combinations of three judges in each subject area. Winer's (1962, pp. 124-132) formula for pooled judge reliability was used in this program. The results indicate

TABLE 4.1
JUDGES' RATINGS
MEANS AND STANDARD DEVIATIONS
SCIENCE ESSAYS

Item #	N		Judge 1	Judge 2	Judge 3	Judge 4
1111	23	$\bar{X}$	4.30	4.22	3.70	3.52
		S.D.	0.88	0.74	0.97	1.24
1112	23	$\bar{X}$	3.87	3.87	3.26	3.13
		S.D.	1.22	1.32	1.32	1.39
1121	18	$\bar{X}$	2.83	2.83	3.44	3.44
		S.D.	1.50	1.38	1.69	1.20
1122	18	$\bar{X}$	3.11	3.24	3.33	3.33
		S.D.	1.37	1.15	1.19	1.19
2111	17	$\bar{X}$	3.53	3.12	3.53	3.24
		S.D.	1.46	1.22	1.07	1.20
2112	17	$\bar{X}$	3.82	3.18	2.94	3.06
		S.D.	1.51	1.24	1.34	1.34
2121	21	$\bar{X}$	2.86	1.85	2.86	3.10
		S.D.	1.39	0.99	1.01	1.26
2122	21	$\bar{X}$	1.81	3.14	3.19	2.33
		S.D.	1.25	1.39	0.98	0.97
3111	15	$\bar{X}$	3.40	4.21	2.80	3.27
		S.D.	1.50	0.70	1.32	1.16
3112	15	$\bar{X}$	2.33	3.53	2.87	3.13
		S.D.	1.50	1.25	1.13	1.25
3121	31	$\bar{X}$	3.61	4.23	3.77	3.74
		S.D.	1.45	0.80	1.15	1.15
3122	31	$\bar{X}$	1.40	3.25	2.60	3.43
		S.D.	0.93	1.04	1.25	1.04
Total	250	$\bar{X}$	3.04	3.44	3.21	3.26
		S.D.	1.57	1.29	1.24	1.22

TABLE 4.2
JUDGES' RATINGS
MEANS AND STANDARD DEVIATIONS
SOCIAL STUDIES ESSAYS

Item #	N		Judge 1	2	3	4
1211	22	$\bar{X}$	2.91	3.23	3.09	3.09
		S.D.	1.19	1.23	1.11	1.41
1212	22	$\bar{X}$	3.95	3.71	2.95	3.59
		S.D.	1.24	1.23	1.59	1.33
1221	26	$\bar{X}$	3.12	2.50	2.50	3.23
		S.D.	0.99	1.02	1.20	0.91
1222	27	$\bar{X}$	3.22	3.44	3.41	3.11
		S.D.	1.37	1.31	1.22	1.40
2211	13	$\bar{X}$	3.62	3.00	3.23	3.75
		S.D.	1.19	1.41	1.17	1.36
2212	10	$\bar{X}$	3.80	2.60	3.00	3.30
		S.D.	1.40	0.97	1.10	1.49
2221	11	$\bar{X}$	2.64	2.45	2.45	3.27
		S.D.	1.03	1.29	1.29	1.01
2222	11	$\bar{X}$	3.80	3.50	3.00	3.45
		S.D.	1.23	1.08	1.18	1.04
3211	19	$\bar{X}$	2.53	2.68	2.84	2.89
		S.D.	1.26	1.00	1.01	0.99
3212	20	$\bar{X}$	2.60	3.00	2.68	2.47
		S.D.	1.35	0.92	1.25	1.22
3221	13	$\bar{X}$	3.31	3.08	3.31	3.46
		S.D.	1.03	1.00	1.25	1.27
3222	13	$\bar{X}$	2.46	2.85	2.62	3.46
		S.D.	1.39	1.14	1.19	1.13
Total	209	$\bar{X}$	3.13	3.04	2.98	3.21
		S.D.	1.29	1.19	1.26	1.24

TABLE 4.3
JUDGES' RATINGS
MEANS AND STANDARD DEVIATIONS
ENGLISH ESSAYS

Item #	N		Judge 1	2	3	4
1311	18	$\bar{X}$	4.27	3.56	3.28	3.89
		S.D.	0.83	1.10	1.32	1.41
1312	18	$\bar{X}$	3.56	2.83	2.94	3.44
		S.D.	1.10	1.42	1.55	1.38
1321	7	$\bar{X}$	3.71	3.29	3.00	2.86
		S.D.	1.60	1.38	1.29	1.57
1322	7	$\bar{X}$	3.43	2.43	2.71	3.29
		S.D.	1.51	1.40	1.38	1.38
2311	18	$\bar{X}$	3.44	2.72	2.83	3.56
		S.D.	1.04	1.07	0.92	0.98
2312	18	$\bar{X}$	3.67	3.56	3.06	3.56
		S.D.	1.14	1.34	1.21	1.20
2321	25	$\bar{X}$	3.08	2.88	2.92	4.20
		S.D.	1.15	1.01	1.22	1.12
2322	25	$\bar{X}$	2.67	2.48	2.52	3.84
		S.D.	1.17	1.19	1.19	1.34
3321	8	$\bar{X}$	4.50	3.25	3.13	3.38
		S.D.	0.76	1.28	1.36	1.41
3322	8	$\bar{X}$	4.13	4.00	3.00	3.00
		S.D.	0.83	0.93	1.20	1.51
3331	21	$\bar{X}$	2.95	2.38	2.71	3.29
		S.D.	0.86	1.12	1.10	1.49
3332	22	$\bar{X}$	2.81	2.55	3.11	4.00
		S.D.	1.10	1.01	1.28	1.38
Total	195	$\bar{X}$	3.37	2.91	2.91	3.65
		S.D.	1.19	1.23	1.23	1.34

TABLE 4.4
JUDGES' RATINGS
MEANS AND STANDARD DEVIATIONS
FOREIGN LANGUAGE ESSAYS

Item #	N		Judge 1	Judge 2	Judge 3	Judge 4
1411	19	$\bar{X}$	2.37	3.05	1.89	2.74
		S.D.	1.01	0.97	1.15	1.15
1412	18	$\bar{X}$	3.33	3.33	3.28	3.18
		S.D.	1.53	1.19	1.60	1.29
1421	19	$\bar{X}$	2.47	2.37	2.37	2.53
		S.D.	1.65	1.64	1.67	1.71
1422	19	$\bar{X}$	3.74	3.58	3.47	4.11
		S.D.	1.19	1.12	1.50	1.20
2411	8	$\bar{X}$	2.50	2.75	2.63	3.00
		S.D.	1.69	1.49	1.41	1.31
2412	8	$\bar{X}$	2.75	2.88	2.00	2.88
		S.D.	1.49	1.46	0.53	1.25
2421	13	$\bar{X}$	3.15	2.85	2.62	3.23
		S.D.	1.46	1.21	1.61	1.69
2422	13	$\bar{X}$	2.23	2.62	2.08	2.54
		S.D.	1.42	1.61	1.26	1.56
3411	6	$\bar{X}$	2.83	3.00	2.00	3.17
		S.D.	1.33	1.41	1.26	1.47
3412	6	$\bar{X}$	3.50	3.17	2.50	3.17
		S.D.	1.52	1.47	1.05	1.47
3421	18	$\bar{X}$	2.44	2.50	1.67	2.78
		S.D.	1.20	1.38	0.97	1.48
3422	16	$\bar{X}$	2.63	2.50	2.69	2.56
		S.D.	1.36	1.67	1.62	1.46
Total	163	$\bar{X}$	2.82	2.88	2.48	2.98
		S.D.	1.43	1.39	1.47	1.46

TABLE 4.5
INTERCORRELATIONS OF ESSAY RATINGS
FOR JUDGE PAIRS
(BY SUBJECT AREA)

Science Essays
N=225

	Judge 2	Judge 3	Judge 4
Judge 1	.505	.580	.595
Judge 2		.453	.420
Judge 3			.677

Social Studies Essays
N=200

	Judge 2	Judge 3	Judge 4
Judge 1	.616	.618	.670
Judge 2		.562	.468
Judge 3			.512

English Essays
N=161

	Judge 2	Judge 3	Judge 4
Judge 1	.518	.535	.398
Judge 2		.448	.400
Judge 3			.553

Foreign Language Essays
N=172

	Judge 2	Judge 3	Judge 4
Judge 1	.755	.817	.608
Judge 2		.672	.519
Judge 3			.652

TABLE 4.6
POOLED RELIABILITY ESTIMATES FOR ALL JUDGES AND ALL COMBINATIONS OF THREE JUDGES USING ANALYSIS OF VARIANCE
(BY SUBJECT AREA)

Subject Area	Judge Code	Pooled Reliability	Adjusted Reliability
Science	0[a]	.813	.817
	1	.755	.758
	2	.815	.817
	3	.739	.747
	4	.746	.754
Social Studies	0	.839	.840
	1	.753	.756
	2	.810	.814
	3	.805	.806
	4	.812	.813
English	0	.750	.785
	1	.664	.723
	2	.705	.748
	3	.657	.703
	4	.732	.753
Foreign Language	0	.883	.889
	1	.811	.825
	2	.858	.870
	3	.831	.832
	4	.891	.899

[a]"Judge code" indicates which judge was eliminated from a particular calculation of the pooled and adjusted reliabilities. Judge code "0" indicates that all judge scores were used in the calculation.

that no judge deviated seriously from the average of the others in his assessments of the student essays, and since the highest reliability estimates were obtained when all four judges' scores in each subject area were used in the calculations, the four-judge average grade for each essay was employed throughout the study as the criterion measure in the multiple regression analyses. The adjusted reliabilities, also reported in Table 4.6, were usually slightly higher than the original pooled reliabilities and eliminate the effect of mean differences among the raters' grades. It is the adjusted reliabilities which have been employed in calculating the attenuated MULTR's which are reported in the summary tables following the descriptions of the results obtained in each analysis series.

Judges' Ratings of Teacher Key and Model Student Essay Effectiveness

The same judges who rated the student essays, generating the criterion measures, also rated the teacher keys and model student essays provided for the essay questions. These ratings were made both before and after the student essays related to each question had been graded. Means and standard deviations of the judges' ratings of teacher key effectiveness are shown in Table 4.7. The highest average rating was given for the teacher keys available for the essay questions in the foreign language area. The lowest

TABLE 4.7
MEANS AND STANDARD DEVIATIONS
TEACHER KEY EVALUATIONS
BEFORE AND AFTER RATING STUDENT ESSAYS
(FOUR-POINT RATING SCALE)

Subject Area	Judge	Before		After	
		Means	S.D.	Means	S.D.
Science	1	1.750	1.055	2.167	1.193
	2	2.083	1.165	2.085	0.900
	3	2.000	1.128	2.250	1.055
	4	2.833	1.030	2.750	1.138
	Average	2.167		2.313	
	Overall Average:	2.240			
Social Studies	1	2.500	1.567	2.583	1.621
	2	2.917	1.240	2.917	1.443
	3	3.083	0.996	3.167	0.937
	4	2.833	1.403	3.083	1.379
	Average	2.833		2.938	
	Overall Average:	2.886			
English	1	3.083	0.669	3.750	0.452
	2	1.833	0.718	1.833	1.030
	3	2.000	0.953	1.833	0.937
	4	2.083	1.084	2.167	1.030
	Average	2.250		2.396	
	Overall Average:	2.323			
Foreign Language	1	1.417	0.515	1.333	0.651
	2	1.167	0.389	1.083	0.289
	3	1.500	0.674	1.500	0.674
	4	1.667	0.651	1.667	0.651
	Average	1.438		1.396	
	Overall Average:	1.417			

ratings were given to the teacher keys for the social studies questions. The model student essays, as described in Table 4.8, generally received lower ratings as aids in the determination of student essay grades. The model student essays in the science, English, and social studies areas maintain the same relative standings on overall effectiveness as do the teacher keys for the questions in these areas; however, in the foreign language area, the model student essays received the lowest average ratings, while the teacher keys for the questions in this area had received the highest average rating of the four subject area groups. In all subject areas except foreign language, the judges tended to have a somewhat lower estimate of both the teacher key and model student essay effectiveness following their grading of the student essays. (A four-point rating scale for teacher key and model student essay effectiveness had been used, with "1" the highest possible rating and "4" the lowest.) Judge agreement as to teacher key and model student essay effectiveness also tended to diminish in the ratings after the grading of the essays in all subject areas except foreign language.

Table 4.9 displays intercorrelations of ratings of teacher key effectiveness for all judge pair combinations, by subject area. The most consistency in rating was demonstrated by the social

TABLE 4.8
MEANS AND STANDARD DEVIATIONS
MODEL STUDENT ESSAY EVALUATIONS
BEFORE AND AFTER RATING STUDENT ESSAYS
(FOUR-POINT RATING SCALE)

Subject Area	Judge	Before		After	
		Means	S.D.	Means	S.D.
Science	1	2.500	1.243	3.250	1.138
	2	2.583	1.311	2.500	1.243
	3	2.583	1.165	2.833	1.030
	4	2.417	1.505	3.083	0.996
	Average	2.521		2.917	
		Overall Average: 2.719			
Social Studies	1	2.750	1.545	2.833	1.528
	2	3.000	1.348	3.083	1.443
	3	3.417	1.084	3.500	0.905
	4	2.833	1.586	3.083	1.443
	Average	3.000		3.125	
		Overall Average: 3.063			
English	1	3.333	1.073	3.833	0.718
	2	2.500	1.446	2.583	1.564
	3	1.917	1.564	2.333	1.614
	4	2.333	1.557	2.417	1.505
	Average	2.521		2.792	
		Overall Average: 2.657			
Foreign Language	1	3.917	1.621	3.917	1.621
	2	3.750	1.865	3.667	1.969
	3	3.750	1.865	3.667	1.969
	4	4.000	1.537	4.000	1.537
	Average	3.854		3.813	
		Overall Average: 3.833			

TABLE 4.9
INTERCORRELATIONS OF JUDGES' EVALUATIONS
OF TEACHER KEY EFFECTIVENESS
BEFORE AND AFTER RATING STUDENT ESSAYS

		Judge	2	3	4
Science		1	.388	.458	.460
	Before	2		.554	.467
		3			.626
		1	-.014	.397	.703
	After	2		.167	.022
		3			.587
Social Studies		1	.725	.786	.703
	Before	2		.963	.670
		3			.661
	After	1	.800	.648	.627
		2		.616	.598
		3			.621
English		1	.032	.000	-.261
	Before	2		.797	.604
		3			.616
		1	-.098	-.322	-.488
	After	2		.439	.114
		3			.596
Foreign Language		1	.076	.655	-.090
	Before	2		-.346	-.120
		3			.414
		1	-.161	.828	.286
	After	2		-.234	-.322
		3			.621

studies rater group (recall that the lowest average ratings of teacher key effectiveness had been given to the keys in this subject area). For the other three subject areas, the judge pair intercorrelations are, generally speaking, discernibly lower than for the social studies raters' intercorrelations. The pairs' ratings tend to become more consistent in the ratings done following the grading of the student essays. The reverse is true for the before-and-after intercorrelations for judge pairs in the ratings of model student essay effectiveness (Table 4.10), though not dramatically so. The judge pair intercorrelations were generally higher and more consistent for the ratings of model student essay effectiveness. The greatest consistency here was found among the foreign language raters, the lowest overall consistency among judge pairs for the science area. Correlations between each judge's "before-and-after" ratings are presented in Table 4.11, which provides ample evidence that the raters' opinions of the teacher key and model student essay effectiveness changed once the student essays associated with each question had been rated.

Judges' Ratings of Essay Question "Content Specificity"

As described in Chapter III, five raters not necessarily possessing a teaching background in the subject area for which they performed ratings, were employed to rate the twelve essay

TABLE 4.10
INTERCORRELATIONS OF JUDGES' EVALUATIONS
OF MODEL STUDENT ESSAY EFFECTIVENESS
BEFORE AND AFTER RATING STUDENT ESSAYS

		Judge	2	3	4
Science		1	.530	.408	.559
	Before	2		.769	.649
		3			.834
		1	.096	.039	.621
	After	2		.639	.477
		3			.724
Social Studies		1	.611	.720	.835
	Before	2		.871	.680
		3			.626
		1	.543	.855	.790
	After	2		.801	.651
		3			.801
English		1	.640	.776	.798
	Before	2		.623	.767
		3			.722
		1	.661	.680	.743
	After	2		.348	.775
		3			.449
Foreign Language		1	.985	.985	.949
	Before	2		1.000	.983
		3			.983
		1	.987	.987	.948
	After	2		1.000	.961
		3			.961

TABLE 4.11
CORRELATIONS BETWEEN "BEFORE" AND "AFTER" RATINGS
TEACHER KEYS AND MODEL STUDENT ESSAYS

Subject Area	Judge 1	2	3	4
Science				
Teacher Keys	.758	.860	.764	.814
Model Student Essays	.803	.586	.467	.824
Social Studies				
Teacher Keys	.948	.910	.860	.806
Model Student Essays	.905	.934	.881	.960
English				
Teacher Keys	.376	.700	.915	.882
Model Student Essays	.787	.904	.840	.944
Foreign Language				
Teacher Keys	.632	.674	.800	.786
Model Student Essays	1.000	.990	.990	1.000

questions in each subject area as to degree of content specificity -- i.e., how much the content of an essay response, as opposed to its style, would appear to require emphasis in the grading of the question. For this purpose, the raters examined the essay question itself and also the accompanying teacher key and model student essay. Two of the raters evaluated all 48 essay items; each of the other sixteen raters evaluated only the twelve essay questions comprising a set for a particular subject area. A ten-point scale was provided for estimating degree of content specificity. Table 4.12 presents the means and standard deviations of content specificity ratings, by rater, for each subject area. With "1" on the scale representing the highest degree of content specificity, and "10" indicating the lowest degree, the reported figures in the table demonstrate that the raters generally considered the group of essay questions for each subject area to be rather content-specific (subject area averages). The highest average content specificity rating was received by the foreign language area essay questions, the lowest by the English items. The most cohesive group in terms of item content specificity would appear to be the foreign language group.

Table 4.13 reveals that rater pair intercorrelations were quite low, overall; the greatest agreement was found among the

TABLE 4.12
MEANS AND STANDARD DEVIATIONS
CONTENT SPECIFICITY RATINGS
(TEN-POINT SCALE)

Subject Area	Judge	Mean	S.D.
Science	1	4.583	1.443
	2	4.750	2.864
	3	3.917	1.311
	4	4.833	1.700
	5	4.000	2.763
	Average:	4.417	
Social Studies	1	4.500	1.624
	2	2.667	1.923
	3	4.250	2.667
	4	6.667	1.923
	5	5.250	1.960
	Average:	4.667	
English	1	4.917	2.021
	2	7.417	1.505
	3	6.333	1.614
	4	4.583	2.466
	5	3.833	2.443
	Average:	5.417	
Foreign Language	1	3.417	1.240
	2	4.167	3.099
	3	3.417	0.669
	4	2.917	1.443
	5	2.583	1.505
	Average:	3.300	

TABLE 4.13
JUDGE INTERCORRELATIONS
CONTENT SPECIFICITY RATINGS

Subject Area	Judge	2	3	4	5
Science	1	-.071	.412	-.217	-.182
	2		.163	.500	.540
	3			.688	.652
	4				.834
Social Studies	1	-.087	.409	-.262	.671
	2		-.461	.066	.193
	3			.301	.578
	4				.217
English	1	-.107	-.325	-.044	.015
	2		.125	.076	.540
	3			.792	.684
	4				.651
Foreign Language	1	.311	-.228	.174	.248
	2		.578	.654	.757
	3			.699	.730
	4				.903

foreign language raters. In Table 4.14, the results of the "cycling" pooled reliability estimates program, when employed for the content specificity raters, are displayed. (Winer's formula was again used, as for the results displayed in Table 4.6.) In contrast to the pooled reliability estimates obtained for the judges of the student essays, the reliabilities described in Table 4.14 change substantially as one or another of the judges' ratings are eliminated from the calculations. Therefore, in this case, the content specificity ratings given by the judge whose elimination resulted in the highest pooled reliability estimate for a subject area group were not utilized when the essays were grouped on the basis of content specificity (see "Analyses by Content Specificity Level" below).

TABLE 4.14
POOLED RELIABILITY ESTIMATES FOR ALL JUDGES
AND ALL COMBINATIONS OF FOUR JUDGES
USING ANALYSIS OF VARIANCE
CONTENT SPECIFICITY RATINGS

Subject Area	Judge Code	Pooled Reliability	Adjusted Reliability
Science	0	.698	.703
	1*	.771	.779
	2	.751	.751
	3	.473	.465
	4	.523	.531
	5	.460	.479
Social Studies	0	.760	.749
	1	.601	.577
	2*	.821	.814
	3	.569	.550
	4	.821	.814
	5	.601	.584
English	0	.730	.712
	1*	.818	.809
	2	.585	.554
	3	.574	.546
	4	.763	.744
	5	.580	.552
Foreign Language	0	.664	.648
	1*	.722	.718
	2	.698	.682
	3	.578	.558
	4	.484	.461
	5	.418	.383

*Denotes judge in each subject area whose ratings were not considered when ranking essay items on the basis of content specificity.

Multiple Regression Analyses

Introduction

Multiple regression analysis is a technique for obtaining the "best fit" of a set of observations of independent and dependent variables by an equation of the form:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

where y is the dependent variable; x_1, x_2 ... are the independent variables; b_0 is a constant (the "intercept" reported in the tables presented below); and b_1, b_2 ... are the coefficients to be determined by the analysis -- an estimate of the relative predictive "strength" of an independent variable, when considered in conjunction with a particular criterion measure, a particular combination of independent variables, and a particular sample. When a stepwise multiple regression analysis is used, a series of intermediate regression equations are obtained by adding one variable at a time to the equation. In the particular procedure used for the analyses presented below, the variable entered at each step in the equation is the one which, of those independent variables not previously entered in the partial regression equation, accounts by itself for the most remaining criterion variance.

The tables presented below provide the following information for each analysis (see TSAR Tele-Storage and Retrieval System

User's Manual, 1970, pp. 41-48, for a description of the procedures used here to carry out the stepwise multiple regression analyses).

1. The name of the variable entered at each step in the analysis

Cumulative Measures

2. The multiple regression coefficient obtained at each step in the analysis

3. The standard error of estimate associated with the multiple regression equation at a given point

4. An F-value giving the significance of the equation at a given point

Final Step Values

5. The final b-weight (coefficient) calculated for each independent variable in the final step of the analysis

6. The standard error of the final b-weight value

7. An F-value indicating the significance of adding a particular predictor to the equation (a test of the null hypothesis that the beta weight, i.e., the regression coefficient of the predictor in the population under investigation, equals zero)

Except for the empirical cross-validation procedures carried out for the multiple regression analyses using all essays (described below), estimated shrinkage of the MULTR -- the depression in the multiple correlation coefficient expected when the derived prediction equation is applied to a new sample of data -- was calculated in all cases using the following formula (McNemar, 1962,

p. 184):

$$R_{shrunken} = \sqrt{1 - (1 - R^2)\left(\frac{N - 1}{N - n}\right)}$$

where: R = original MULTR value obtained

N = sample size

n = number of predictor variables employed

Because of the lack of perfect reliability of the criterion measure, these shrunken MULTR's were then corrected for attenuation, to estimate the predictive value of the derived equations had a perfectly reliable criterion been employed. The following formula was used to estimate the attenuated MULTR (McNemar, 1962, p. 153):

$$r_{tt} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

where: r_{tt} = attenuated MULTR, defined as the correlation of the underlying true scores

r_{xy} = shrunken MULTR

r_{xx} = reliability of the predictor variables

r_{yy} = reliability of the criterion measure

Here, r_{xx} has been taken to be 1.00, since it represents the reliability of the computer-generated essay grades; these values are perfectly reliable since, given the same methods for computing the

scores on each variable a number of times, the computer would always produce the same values for the variables.

The significance of the shrunken MULTR's (the shrunken MULTR being the most conservative of the three MULTR values obtained in every case) was calculated using the following formula:

$$F = \frac{(R^2)\ df}{(1 - R)^2 n}$$

where: R = shrunken MULTR

n = number of predictors

df = N - (n + 1)

The remainder of this chapter is devoted to a description of the results obtained in four series of analyses carried out using the sample of student essays; all of these analyses employed multiple regression analysis as the major hypothesis-testing tool.

Analyses Using All Essays

Two intercorrelation matrices are presented to introduce this series of analyses. In both, the intercorrelations of all predictors and the criterion measure are displayed. Table 4.15 includes intercorrelations for the content variables with all other measures when the content variable values were determined by a procedure in which the entire student essay was used as the "search unit". In Table 4.16, the intercorrelations for the content variables with the other measures are somewhat different because the content-related predictors here were determined using the essay sentence as the search unit for values for the content variables. The distinctions between these two procedures will be described in detail below.

In examining these matrices, it should be kept in mind that the signs of the correlations between predictors and the criterion should be reversed, since "1" was the highest possible grade and "5" was the lowest. It is apparent from the matrices that the length of essay variable is the most highly correlated with the criterion. Following length of essay, the two predictors most highly related to the criterion measure were "teacher key phrases partially matched" (KEYP) and "model student

TABLE 4.15
INTERCORRELATION MATRIX
ALL ESSAYS
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

	SPELL	PREP	PROPN	MEANW	SDW	MEANS	SDS	COMMA	DASH	RELPR	KEYC	KEYP	MODC	MODP	OTHP	GRADE
LENGT	-098	115	007	-101	-078	374	344	209	-003	107	192	303	197	279	-075	-447
SPELL		-062	-035	090	052	037	014	-048	027	-016	-035	-083	-045	-090	043	133
PREP			-072	-220	-048	234	033	-035	-031	-107	-020	073	-045	075	-010	-067
PROPN				151	070	-068	050	062	047	-074	015	018	008	-024	-004	014
MEANW					636	-184	-082	191	285	-088	032	112	064	095	-017	-105
SDW						-152	-047	070	414	-123	077	041	095	043	-037	-104
MEANS							129	211	-074	154	137	137	034	065	074	-146
SDS								033	-059	114	035	067	052	014	023	-112
COMMA									-012	041	099	067	096	070	-047	-184
DASH										-004	042	019	064	033	004	-020
RELPR											025	-031	017	-049	016	-054
KEYC												-254	553	-131	-061	-233
KEYP													-106	634	006	-358
MODC														-211	-073	-241
MODP															006	-306
OTHP																-104

N=758.
All entries multiplied by 1000.

TABLE 4.16
INTERCORRELATION MATRIX
ALL ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

	SPELL	PREP	PROPN	MEANW	SDW	MEANS	SDS	COMMA	DASH	RELPR	KEYC	KEYP	MODC	MODP	OTHP	GRADE
LENGT	-058	115	008	-101	-078	375	344	203	-003	107	135	312	098	234	030	-447
SPELL		-062	-035	090	052	037	014	-048	028	-016	-010	-072	-021	-079	017	133
PREP			-072	-220	-048	234	033	-095	-031	-107	-003	051	-043	038	-026	-017
PROPN				151	070	-068	050	062	047	-074	038	012	015	-010	048	014
MEANW					635	-184	-082	101	285	-088	055	125	078	117	-012	-105
SDW						-152	-087	073	414	-123	093	051	115	059	-077	-104
MEANS							129	211	-074	154	123	179	063	094	051	-146
SDS								033	-059	114	007	082	031	042	123	-112
COMMA									-012	041	064	079	090	068	043	-184
DASH										-004	037	030	073	044	-002	-020
RELPR											-001	-029	-099	-042	067	-054
KEYC												-182	453	-097	-028	-197
KEYP													-024	626	-019	-378
MODC														-052	011	-205
MODP															-023	-350
OTHP																000

N=758.

All entries multiplied by 1000.

essay phrases partially matched" (MODP). Four out of the five content-related predictor variables yield among the highest correlations with the criterion measure. Only OTHP ("phrases from other sources partially matched") is among the less highly correlated predictors.

Note that length of essay is substantially correlated with a number of the other predictors, especially "mean sentence length" (MEANS), "standard deviation of sentence length" (SDS), "teacher key phrases partially matched" (KEYP), and "model student essay phrases partially matched" (MODP). High correlations are shown to exist between the variables representing complete matches on teacher keys and model student essays, and between the variables representing partial matches on these sources.

A total of five analyses were carried out using the entire available sample of 758 essays. First, only the eleven stylistic variables were allowed to enter into a multiple regression equation to predict the criterion values. In this analysis, the results of which are reported in Table 4.17, four significant predictors were discerned -- "length of essay" (LENGT), "mean word length" (MEANW), "proportion of spelling errors" (SPELL), and "standard deviation of word length" (SDW). Following the inclusion of the first five variables, the addition of the remaining six predictors raised the

TABLE 4.17
STEPWISE MULTIPLE REGRESSION
ALL ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0520	.0042	156.9940**	.4469	.9971	197.4897**
2	MEANW	-.0118	.0049	5.6963*	.4717	.9834	113.0623**
3	SPELL	.0115	.0036	9.9587**	.4827	.9774	79.9082**
4	COMMA	-.0072	.0041	3.1284	.4859	.9760	60.9130**
5	SDW	-.0111	.0049	5.0683*	.4895	.9744	49.6210**
6	PROPN	.0058	.0044	1.7568	.4916	.9737	41.7633**
7	DASH	.0061	.0047	1.7412	.4931	.9734	36.0341**
8	SDS	.0041	.0040	1.0231	.4940	.9734	31.6383**
9	RELPR	-.0027	.0040	.4675	.4945	.9737	28.1644**
10	PREP	.0010	.0038	.0686	.4946	.9743	25.3263**
11	MEANS	.0003	.0004	.0005	.4946	.9749	22.9951**

Intercept: 3.0788

* Significant at .05 level
** Significant at .01 level

MULTR only about .005. The final MULTR achieved was .4946. The great importance of the LENGT variable in the prediction equation was the major feature of almost all of the analyses conducted with the essays. As will be demonstrated in the results described below, this variable maintained its prominence in the prediction equation even when the content-related predictors were included in the calculations.

In order to evaluate meaningfully the results of the next two analyses done using the entire group of essays, an understanding of the distinction between the "essay search unit" and the "sentence search unit" is necessary. As described in Chapter III, a set of "computer-usable key phrases" had been prepared for each essay question, by the author, using both the teacher key and the model student essay (where available) for the question as sources for these phrases. Each phrase was coded as to its source (whether teacher key or model essay). If the same thought had been included in <u>both</u> the teacher key and model essay, the computer-usable key phrase conveying that thought was coded for both sources. In a few cases, the author felt that an ideal response to an essay item should cover information not contained in either the teacher key or the model essay for that item. In these cases, an additional phrase (or phrases) representing this information

was prepared for inclusion among the computer key phrases for the item. These phrases were coded as being from a source other than the teacher key or model essay.

The Stage 4 computer program prepared for the AEC-II study took the student essays for each question and the list of computer-usable key phrases for the question and searched through each student essay to find "matches" in the essay for the elements (words or "idioms") appearing in the key phrases. A count was made of each phrase element matched. Complete matches on key phrases were distinguished from partial matches; the source code for the key phrase, in combination with the identification of a complete or partial match, determined the specific content variable to which the "match" count would be applied. When all matches within the student essay had been counted, the six separate scores thus obtained (complete and partial matches on phrases representing each of the three sources) would each be divided by the sum of the elements contained in the phrases of the appropriate source code. For example, consider an essay question which has associated with it four computer-key phrases coded as having the teacher key as their "source". These phrases are varying lengths -- l_1, l_2, l_3, and l_4. If, in a particular student essay, the first phrase was completely matched, and the remaining three phrases were each

partially matched (the number of elements in each phrase located in the essay being p_2, p_3, and p_4, with $p_2 < l_2$, $p_3 < l_3$, and $p_4 < l_4$), the resulting scores for that essay on the KEYC and KEYP variables would be as follows:

$$\text{KEYC} = \frac{l_1}{l_1 + l_2 + l_3 + l_4}$$

$$\text{KEYP} = \frac{p_2 + p_3 + p_4}{l_1 + l_2 + l_3 + l_4}$$

These "variable scores," after being transformed ($\bar{X}$=50.0, s.d. =10.0) to standardize each variable, were then employed in the multiple regression equations.

In addition, two different methods for searching out matches on the content variables were used. In the analyses employing the first method, the student essay as a whole was used in obtaining matches on the words contained in a computer-key phrase. With this procedure, a student essay could achieve a complete match with a key phrase even though the words in the key phrase had each been located in a different sentence in the student essay. This procedure essentially reduced the problem of determining matches on the key phrases to a search for each key phrase element, and took no account of the contingencies of key phrase words in the student text.

The second search procedure was more stringent. In the analyses employing this method, each sentence of the student text was compared once with each computer key phrase and the highest number of matches received for a particular key phrase on a sentence in a particular text determined the phrase score for that phrase assigned to that student text, thus increasing the content variable values for those essays in which key word contingencies more closely resembled the element groups which comprised the computer key phrases.

In the second analysis carried out using all 758 essays as a group, the five content-related variables were included among the predictors and the essay was used as the "search unit" for the matches on the key phrases which determined the values of the content variables. (Recall that one of the six original content variables -- OTHC, "other source phrases completely matched" -- was excluded from all analyses because such small values were obtained for this variable that intercorrelation measures between this variable and all of the others were invariably zero, which prevented the computer program used from carrying out the multiple regression solution when this variable was included among the predictors.) In this analysis (Table 4.18), four of the five content variables emerged as significant predictors, with the teacher key-related

TABLE 4.18
STEPWISE MULTIPLE REGRESSION
ALL ESSAYS
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0322	.0043	57.2460**	.4469	.9971	197.4897**
2	KEYP	-.0311	.0049	40.2303**	.5041	.9632	134.6264**
3	KEYC	-.0247	.0051	23.2000**	.5589	.9255	119.4876**
4	SEN	-.0092	.0046	4.0305*	.5668	.9200	93.2581**
5	MODC	-.0165	.0053	9.8179**	.5727	.9160	76.8277**
6	MODP	-.0134	.0053	6.4197*	.5784	.9121	65.8690**
7	SPELL	.0075	.0034	4.8831*	.5823	.9096	57.5326**
8	COMMA	-.0069	.0038	3.3031	.5856	.9075	51.1653**
9	OTHP	.0126	.0067	3.5242	.5884	.9058	46.0900**
10	RELPR	-.0051	.0037	1.8967	.5898	.9053	41.7198**
11	DASH	.0061	.0043	1.9907	.5911	.9048	38.1384**
12	PROPN	.0050	.0041	1.5000	.5919	.9047	35.0535**
13	MEANW	-.0044	.0047	.8883	.5928	.9045	32.4832**
14	MEANS	.0027	.0039	.4864	.5933	.9047	30.1965**
15	PRET	.0015	.0036	.1834	.5934	.9052	28.1660**
16	SDS	.0012	.0038	.1088	.5935	.9057	26.3821**

N=758.
Intercept: 3.0788

* Significant at .05 level
** Significant at .01 level

variables much more prominent in their effectiveness than the model student essay-related variables. Only the content variable, "phrases from other sources partially matched," was not significant. The LENGT variable remained the single most efficient predictor among the variables; as in the previous analysis, the standard deviation of word length and the proportion of spelling errors were significant. Mean word length, however, was no longer a significant predictor when the content variables were employed. The MULTR attained for all essays as a group when all sixteen independent variables were used in the predictive equation and the essay was employed as the search unit for matches on the content variables was .5935.

When the sentence was used as the search unit for matches on the content variables, a somewhat different pattern was observed in the multiple regression analysis results (Table 4.19). All predictors shown to be significant in the previous analysis were significant here as well and were entered into the equation in roughly the same order -- the exceptions being that the two model essay-related variables were entered here before standard deviation of word length. A new significant predictor emerged, the proportion of commas; since this variable was not a significant predictor when only stylistic variables were employed

TABLE 4.19
STEPWISE MULTIPLE REGRESSION
ALL ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0364	.0041	78.1909**	.4469	.9971	197.4897**
2	KEYP	-.0269	.0047	32.2633**	.5121	.9577	140.7359**
3	KEYC	-.0221	.0049	20.2999**	.5513	.9311	114.8834**
4	MODP	-.0179	.0051	12.3968**	.5607	.9247	90.3317**
5	MODC	-.0167	.0050	11.2579**	.5713	.9171	76.2873**
6	SDW	-.0102	.0046	4.8849*	.5777	.9126	65.6477**
7	SPELL	.0083	.0034	6.0540*	.5828	.9091	57.7117**
8	COMMA	-.0081	.0038	4.601[illegible]	.5864	.9068	51.3852**
9	MEANS	.0063	.0039	2.5882	.5881	.9061	46.0034**
10	RELPR	-.0062	.0037	2.7590	.5900	.9051	41.7742**
11	DASH	.0070	.0043	2.5880	.5918	.9042	38.2803**
12	PROPN	.0054	.0041	1.7226	.5930	.9038	35.2563**
13	MEANW	-.0033	.0047	.4926	.5934	.9041	32.5711**
14	SDS	.0023	.0038	4.6018*	.5936	.9044	30.2502**
15	OTHP	.0014	.0068	.0449	.5937	.9050	28.2018**
16	PREP	.0005	.0036	.0166	.5937	.9055	26.4068**

N=758.
Intercept: 3.0788

* Significant at .05 level
** Significant at .01 level

(Table 4.17), it appears that the COMMA variable interacted with one or more of the content variables, especially when the content variable values were determined using the sentence as search unit. The significance of the standard deviation of sentence length (SDS), combined with its late inclusion into the regression equation, identifies it as a suppressor variable. A final prediction coefficient of .5937 was obtained in this analysis, compared to a MULTR value of .5935 achieved in the corresponding analysis (Table 4.18) in which the essay as a whole had been employed as the search unit for values on the content variables.

Two further analyses using all essays were carried out, to determine if reducing the five content variables to two measures ("phrases partially matched" and "phrases completely matched") would greatly affect the prediction level attained. Table 4.20 presents the results of the first of these analyses, in which the essay was used as the search unit. For these analyses, the "variable scores" obtained using the procedures described above were accumulated into two sums (KEYC + MODC, and KEYP + MODP + OTHP). The two new content variables (CONTC and CONTP) are seen to be highly effective predictors, especially the measure of partially matched phrases. Standard deviation of word length and the proportion of spelling errors were again sig-

TABLE 4.20
STEPWISE MULTIPLE REGRESSION
ALL ESSAYS
STYLISTIC AND COMBINED CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0378	.0042	59.6147**	.4469	.9971	197.4897**
2	CONTP	-.0226	.0023	93.2711**	.5058	.9621	135.8584**
3	CONTC	-.0208	.0025	66.9690**	.5499	.9170	126.5338**
4	SDW	-.0095	.0046	4.2516**	.5759	.9129	97.7921**
5	SPELL	.0077	.0034	5.0610**	.5797	.9105	79.6920**
6	COMMA	-.0069	.0038	3.3118**	.5831	.9084	67.4918**
7	DASH	.0065	.0043	2.2765	.5844	.9079	58.1911**
8	RELPR	-.0052	.0037	1.9061	.5859	.9072	51.2494**
9	MEANW	-.0046	.0047	.9567	.5868	.9072	45.6902**
10	PROPN	.0046	.0041	1.2377	.5876	.9070	41.2504**
11	MEANS	.0024	.0038	.3968	.5880	.9073	37.5279**
12	PREP	.0015	.0036	.1645	.5881	.9078	34.3774**
13	SDS	.0013	.0038	.1132	.5882	.9083	31.7056**

N=758.
Intercept: 3.0788

* Significant at .05 level
** Significant at .01 level

nificant, as they had been in all previous analyses; the proportion of commas was also significant, as in the immediately preceding analysis (Table 4.19). The final MULTR value was .5882, just slightly lower than the value obtained in the corresponding analysis (Table 4.18) in which the five content variables were included separately in the regression equation.

The second analysis carried out using the two combined content variables (Table 4.21), in which the sentence was used as the search unit for the content variables, produced results very similar to those of the analysis described immediately above. The same significant predictors were observed and these predictors had the same relative prominence as in the preceding analysis. The MULTR for this test, .5927, was slightly higher than the value obtained for the analysis which used the essay as the search unit. This distinction was the same between the MULTR's for the different search units when the five content variables had not been combined (Tables 4.18, 4.19).

To determine the estimated shrinkage of the MULTR's, for the analyses using all essays, if essay grades were predicted for a new sample from the same population employing the prediction equations generated for <u>this</u> sample, an empirical cross-validation procedure was carried out. The entire group of 758

TABLE 4.21
STEPWISE MULTIPLE REGRESSION
ALL ESSAYS
STYLISTIC AND COMBINED CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0364	.0041	78.9824**	.4469	.9971	197.4897**
2	CONTP	-.0225	.0023	98.7033**	.5220	.9513	147.9737**
3	CONTC	-.0194	.0026	54.8639**	.5708	.9163	127.1384**
4	SDN	-.0102	.0046	4.9252*	.5772	.9119	98.4239**
5	SPELL	.0083	.0034	6.0712*	.5822	.9085	80.7356**
6	COMMA	-.0079	.0038	4.4312*	.5858	.9062	68.4572**
7	RELPR	-.0062	.0037	2.7395	.5874	.9055	59.0863**
8	DASH	.0071	.0043	2.7113	.5891	.9046	52.1056**
9	MEANS	.0057	.0029	2.2154	.5909	.9037	46.6981**
10	PPOPN	.0052	.0041	1.6272	.5920	.9034	42.2061**
11	MEANW	-.0034	.0047	.5404	.5925	.9036	38.4096**
12	SDS	.0023	.0038	.3900	.5927	.9040	35.2137**
13	PREP	.0004	.0036	.0128	.5927	.9046	32.4648**

N=758.

Intercept: 3.0788

* Significant at .05 level
** Significant at .01 level

essays was divided into thirds, using the student number for each essay response to sort the essays into three subgroups. Since these student numbers had originally been assigned in a completely arbitrary manner, it was assumed that division on this basis would result in a random number of student essays falling into each group (Table 4.22). The data from each combination of two groups were then used to generate MULTR's, using first the stylistic variables alone and then the combination of stylistic and content variables. The procedure was done once with the entire essay as search unit and then with each sentence as search unit for values on the content variables. The b-weights from the equation formed from each combination of two groups were then applied to the data from the group not in the pair in order to generate predicted grades. These predicted grades and the criterion values were then correlated to obtain a new, cross-validated MULTR. The three MULTR's found in this fashion for each of the three sets of variables were then averaged; this average value was attenuated. As shown in Table 4.23, the average cross-validated MULTR obtained for each of the three sets of predictors was significant. The average amount of shrinkage is the same for all three variable sets and there is very little difference in the results whether the essay or the sentence search unit is used.

TABLE 4.22
GROUPING OF STUDENT ESSAYS FOR
CROSS-VALIDATION PROCEDURE

Student Numbers	Group Placement
1 - 3	A
4 - 6	B
7 - 9	C
10 - 12	B
13 - 15	A
16 - 18	C
19 - 21	B
22 - 24	C
25 and above	A

TABLE 4.23
CROSS-VALIDATION RESULTS

Developmental Sample	Original Mult-R	Cross-Validation Sample		Shrunken Mult-R
		Stylistic Variables Only		
Groups A and B	.4987	C		.4546
Groups A and C	.5108	B		.4353
Groups B and C	.4960	A		.4646
			Average	.4514**
			Attenuated	.4947
		Stylistic and Content Variables (Essay as Unit)		
Groups A and B	.5950	C		.5456
Groups A and C	.6067	B		.5450
Groups B and C	.5998	A		.5650
			Average	.5519**
			Attenuated	.6048
		Stylistic and Content Variables (Sentence as Unit)		
Groups A and B	.6038	C		.5480
Groups A and C	.6097	B		.5436
Groups B and C	.5874	A		.5896
			Average	.5604**
			Attenuated	.6141

Groups A and B: N=503
Groups A and C: N=498
Groups B and C: N=507

** Significant at .01 level

In Table 4.24, the series of analyses using all essays are summarized. Estimates of the significance of the difference between the "stylistic only" MULTR and (1) the combination of stylistic and content variables with the essay as search unit, (2) the combination of stylistic and content variables with the sentence as search unit (the F-values reported in the table) are given. The differences are seen to be highly significant.

The multiple regression coefficients obtained when the five content variables were reduced to two sums and employed in the regression equation with the stylistic variables were similar to those obtained when the five content variables had been included separately. However, these MULTR's were not compared directly with the "stylistic only" MULTR since shrinkage values for these two coefficients reported in the table had been estimated mathematically, while the estimated shrinkage had been determined empirically (and thus more conservatively) for the MULTR representing the stylistic analysis of the essays.

TABLE 4.24
SUMMARY OF ANALYSES USING ALL ESSAYS

Variables Employed	Mult-R's Original	Shrunken	Attenuated
Stylistic Only	.4946	(1).4514**	.4947
Stylistic and Content (Essay as Unit)	.5935	(2).5519**	.6048
Significance of difference: F = 22.4444** (1) vs. (2)			
Stylistic and Content (Sentence as Unit)	.5937	(3).5604	.6141
Significance of difference: F = 24.5556** (1) vs. (3)			
Stylistic and Combined Content Variables (Essay as Unit)	.5882	(4).5774**	.6327
Stylistic and Combined Content Variables (Sentence as Unit)	.5927	(5).5820**	.6378

** Significant at .01 level

Analyses by Subject Area

The results of the multiple regression analyses of the essays by subject area are introduced with a series of tables showing the intercorrelations of all variables, both stylistic and content-related, by subject area (Tables 4.25 through 4.28). The variable MODC ("model student essay phrases completely matched") has been eliminated from the analyses of the foreign language area essays, because the number of matches obtained on this variable were so small for this subject area group that zero correlations resulted between this predictor and all other predictors, which prevented the multiple regression program from running to completion when the MODC variable was included among the predictors.

An examination of Table 4.25 reveals a number of the correlation patterns which hold across the four subject areas. As was the case when all essays were analyzed together, comparatively high correlations were found between the variables "mean word length" (MEAN) and "standard deviation of word length" (SDW). For all subject areas considered, the correlation for the "teacher key phrases completely matched" (KEYC) - "model student essay phrases completely matched" (MODC) pair (except, of course, for the foreign language area group, in which MODC did not appear)

TABLE 4.25

INTERCORRELATION MATRIX
SCIENCE ESSAYS
STYLISTIC AND CONTENT VARIABLES

	SPELL	PREP	PROPN	MEANW	SDW	MEANS	SDS	COMMA	DASH	RELPR	KEYC	KEYP	MODC	MODP	OTHP	GRADE
LENGT	-004	192	057	013	-009	406	426	232	070	223	201	373	241	355	-081	-538
SPELL		-045	-061	163	206	049	008	108	067	084	052	-073	-028	-064	007	129
PREP			-049	-086	164	327	110	-084	005	017	-002	185	009	127	029	-079
PROPN				109	-027	086	018	111	018	016	100	-015	099	-105	025	-038
MEANW					599	-051	-011	258	346	-034	145	086	098	163	000	-174
SDW						-045	-018	023	547	-110	142	-036	098	046	-034	-110
MEANS							026	222	-074	138	203	101	070	190	050	-201
SDS								019	-014	143	-023	155	029	065	025	-224
COMMA									-023	003	139	060	119	136	022	-256
DASH										047	155	-003	231	044	-003	-094
RELPR											037	025	050	-015	-050	-017
KEYC												-156	578	-106	071	-191
KEYP													-068	766	029	-441
MODC														-160	-115	-285
MODP															045	-429
OTHP																055

N=225.
All entries multiplied by 1000.

TABLE 4.25

INTERCORRELATION MATRIX
SOCIAL STUDIES ESSAYS
STYLISTIC AND CONTENT VARIABLES

	SPELL	PREP	PROPN	MEANW	SDW	MEANS	SDS	COMMA	DASH	RELPR	KEYC	KEYP	MODC	MODP	OTHP	GRADE
LENGT	-136	024	-002	-190	-105	179	330	174	-055	089	182	308	146	324	-145	-574
SPELL		001	062	046	008	102	047	-118	054	-102	-094	-102	-021	-110	079	121
PREP			-005	-368	-191	131	154	-063	-141	-085	-026	-036	-129	025	-082	055
PROPN				237	168	021	115	009	051	-067	-088	-058	-046	-109	-041	133
MEANW					682	-241	-178	215	333	-044	-002	089	085	071	-040	047
SDW						-106	-136	180	468	-125	131	058	188	-009	-042	-016
MEANS							259	237	-117	114	132	-055	038	-077	135	-026
SDS								-002	-095	085	030	059	078	-060	062	-131
COMMA									-011	111	158	-003	106	077	-084	-136
DASH										-060	031	008	-007	025	-001	078
RELPR											011	017	-070	070	045	-007
KEYC												-231	671	-156	-184	-287
KEYP													-146	687	016	-315
MODC														-203	-116	-238
MODP															-029	-345
OTHP																230

N=200.
All entries multiplied by 1000.

TABLE 4.27

INTERCORRELATION MATRIX
ENGLISH ESSAYS
STYLISTIC AND CONTENT VARIABLES

	SPELL	PREP	PROPN	MEANW	SDW	MEANS	SDS	COMMA	DASH	RELPR	KEYC	KEYP	MODC	MODP	OTHP	GRADE
LENGT	-189	068	-043	-021	086	105	265	123	-008	-060	267	106	257	276	043	-331
SPELL		-059	002	132	022	035	-009	-025	-075	-075	-049	-057	-102	-171	030	185
PREP			-125	-160	-022	082	-062	-212	045	-226	-031	-021	-028	076	067	103
PROPN				183	142	-207	-121	129	079	-233	210	081	025	109	019	-191
MEANW					631	-199	-127	231	117	-259	-024	138	071	016	-053	-157
SDW						-152	-026	170	227	-198	-033	179	111	080	-055	-187
MEANS							295	050	010	268	-008	051	-071	032	105	052
SDS								090	-030	094	177	-041	073	029	-034	-020
COMMA									036	-041	014	098	075	016	-131	-171
DASH										031	-006	-005	028	-039	021	-079
RELPR											-067	-114	018	-208	147	-118
KEYC												-356	416	-061	022	-288
KEYP													-043	464	-080	-125
MODC														-216	018	-379
MODP															041	-178
OTHP																091

N=161.
All entries multiplied by 1000.

TABLE 4.28

INTERCORRELATION MATRIX
FOREIGN LANGUAGE ESSAYS
STYLISTIC AND CONTENT VARIABLES

	SPELL	PREP	PROPN	MEANW	SDW	MEANS	SDS	COMMA	DASH	RELPR	KEYC	KEYP	MODC[a]	MODP	OTHP	GRADE
LENGT	-131	212	-105	-217	-241	790	274	316	073	084	055	461		197	-060	-361
SPELL		-173	-252	-003	-086	-071	-058	-065	027	193	-056	-142		-022	035	158
PREP			-158	-276	-194	393	-125	017	-010	-227	024	144		075	069	-127
PROPN				125	045	-211	045	-003	146	031	-166	071		-070	032	109
MEANW					612	-259	-034	-014	265	014	-028	094		018	025	-110
SDW						-304	022	-106	280	-044	086	-076		-079	-016	-130
MEANS							-103	226	002	016	129	450		256	-090	-343
SDS								123	-150	162	-053	068		000	033	-038
COMMA									-072	-008	-036	173		000	-077	-202
DASH										007	-027	008		000	-061	001
RELPR											102	-070		000	-092	-059
KEYC												-288		000	-300	-255
KEYP														511	076	-511
MODC																
MODP															000	-364
OTHP																097

N=172.
All entries multiplied by 1000.
[a] "MODC" variable eliminated from analysis of foreign language essays. See text for explanation.

and the "teacher key phrases partially matched" (KEYP) - "model student essay phrases partially matched" (MODP) pair are among the highest correlations found in the matrices. Another consistency is the relatively high correlations obtained between the criterion (GRADE) and the "length of essay" (LENGT) variable. Discounting the essay length variable, the highest correlations with the criterion measure are to be found among the content variables for all four subject area groups: in the English essays group and in the foreign language essays group, certain content variables have a higher correlation with the criterion than the correlation between the criterion and the essay length variable.

The science essays and the foreign language essays groups (Tables 4.25 and 4.28 respectively) are similar in that they consist of relatively short responses (compared with the other two subject areas) and the intercorrelation matrices for these two groups share several features. In both cases, and in contrast to the English and social studies groups, the correlation between essay length and mean sentence length is quite high -- a result of the fact that so many of the responses in these two groups consisted of a single sentence (or phrase). Another common feature of these two groups is the fact that the highest correla-

tions with the criterion are obtained for essay length, and the counts of partially matched phrases on both the teacher key and the model student essay. The English essays (Table 4.27) are distinguished by the relatively low correlation of "LENGT" with "KEYP" compared to the other subject area groups.

The first stage in the analyses of the essays by subject area consisted of determining MULTR values for each subject area using only the eleven stylistic variables as predictors. Table 4.29 presents the results of this analysis for the science essays. Only two of the predictor variables were significant -- length of essay and proportion of spelling errors, the first and third variables entered into the equation. The LENGT variable by far accounts for the most variance among the essay grades in this group. Note that the inclusion of the first three variables in the equation yields a MULTR of .5846; the addition of the remaining eight variables as predictors raises the MULTR value only by about .02. The significance of the length of essay variable is a pattern which holds for three of the four subject areas considered, when stylistic variables alone are used as predictors. For all four subject areas, LENGT is the first variable entered; only for the foreign language essay group is *essay length not a significant predictor -- a probable result*

TABLE 4.29

STEPWISE MULTIPLE REGRESSION
SCIENCE ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0590	.0072	58.2928**	.5382	.9119	90.9746**
2	MEANW	-.0139	.0080	2.9745	.5534	.8959	51.6290**
3	SPELL	.0180	.0063	8.0601**	.5846	.8818	38.2616**
4	RELPR	.0130	.0064	2.4079	.5924	.8776	29.7488**
5	COMMA	-.0108	.0065	2.6156	.5984	.8747	24.4387**
6	SDW	-.0070	.0090	.6086	.6000	.8754	20.4449**
7	PROPN	.0050	.0100	.2075	.6005	.8770	17.4898**
8	PREP	.0025	.0064	.1450	.6009	.8787	15.2620**
9	SDS	-.0015	.0068	.0484	.6010	.8807	13.5124**
10	DASH	.0006	.0081	.0063	.6010	.8827	12.1056**
11	MEANS	-.0001	.0072	.0002	.6010	.8848	10.9537**

N=225.
Intercept: 3.2980

** Significant at .01 level

of the uniform brevity of most of the responses in this area. The science essays were unique among the four groups in having the proportion of spelling errors as a significant predictor variable.

When stylistic variables alone were used in an analysis of the social studies essays (Table 4.30), a MULTR of .6145 was obtained -- almost the same value as was demonstrated for the science essays using stylistic variables only as predictors. Length of essay was again the best single predictor; the only other significant variable was "proportion of proper nouns," a stylistic variable which had not been employed in the earlier essay analysis studies done by Page and Paulus, and others.

Three significant predictor variables emerged in the stylistic analysis of the English essay group: length of essay, proportion of relative pronouns (RELPR), and (as was true for the social studies essays) the proportion of proper nouns used (Table 4.31). In contrast to the previous two subject area groups, the contribution of the LENGT variable to the MULTR value was less valuable here, reflected in the lower correlations between this variable and the other stylistic predictors for the English essay group (see Table 4.27). The effectiveness of the proportion of proper nouns predictor in the English and social

TABLE 4.30

STEPWISE MULTIPLE REGRESSION
SOCIAL STUDIES ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0637	.0068	86.7703**	.5741	.8421	97.3546**
2	MEANW	-.0019	.0096	.0415	.5891	.8332	52.3518**
3	SPELL	.0024	.0062	.1529	.5975	.8289	36.2817**
4	RELPR	.0051	.0062	.6781	.6054	.8249	28.2071**
5	COMMA	-.0024	.0068	.1240	.6098	.8235	22.9710**
6	SDW	-.0123	.0090	1.8640	.6117	.8241	19.2395**
7	PROPN	.0182	.0073	6.1033*	.6132	.8250	16.5370**
8	PREP	.0052	.0067	.6120	.6139	.8266	14.4420**
9	SDS	.0012	.0067	.0328	.6143	.8285	12.7970**
10	DASH	.0120	.0068	3.0999	.6144	.8306	11.4634**
11	MEANS	.0065	.0068	.8913	.6145	.8327	10.3710**

N=200.
Intercept: 3.0987

* Significant at .05 level
** Significant at .01 level

TABLE 4.31

STEPWISE MULTIPLE REGRESSION
ENGLISH ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0331	.0074	19.5163**	.3814	.9078	19.6202**
2	MEANW	-.0127	.0096	1.4839	.3901	.8888	14.1842**
3	SPELL	.0117	.0074	2.4904	.4351	.8719	12.2236**
4	RELPR	-.0236	.0079	8.8595**	.4706	.8571	11.0948**
5	COMMA	-.0072	.0074	.9246	.4875	.8509	9.6683**
6	SEN	-.0075	.0092	.6602	.4915	.8514	8.1787**
7	PROPN	-.0206	.0078	6.8677**	.4966	.8514	7.1555**
8	PREP	.0014	.0077	.0332	.5000	.8522	6.3362**
9	SDS	.0039	.0076	.2646	.5013	.8543	5.6330**
10	DASH	-.0017	.0083	.0436	.5015	.8571	5.0413**
11	MEANS	.0063	.0077	.6714	.5017	.8598	4.5565**

N=161.
Intercept: 3.1754

* Significant at .05 level
** Significant at .01 level

studies groups is not surprising when one considers the nature of a typical "recall" essay question in these two areas at the junior high and high school level. The emergence of the proportion of relative pronouns predictor is a distinguishing feature of the English essay group.

The foreign language essay group was unique among the four subject areas in that the length of essay was not a significant predictor (Table 4.32), although this variable, as for the other three subject areas, accounts by itself for more criterion variance than any of the other stylistic variables. The single significant stylistic predictor for the foreign language essay group was "standard deviation of word length" (SDW), and only for this group was SDW a significant predictor. The MULTR resulting from the inclusion of the eleven stylistic predictors was .4988, a value very close to that achieved for the English essays and substantially lower than the MULTR values obtained for the science and social studies essays using the stylistic variables alone as predictors.

In a second series of analyses by subject area, multiple regression coefficients were obtained for each of the four subject area groups using both content and stylistic predictors -- a total of sixteen independent variables, eleven stylistic

TABLE 4.32

STEPWISE MULTIPLE REGRESSION
FOREIGN LANGUAGE ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0367	.0199	3.4017	.3609	1.2118	25.4695**
2	SDW	-.0284	.0123	5.3260*	.4249	1.1797	18.6239**
3	MEANS	-.0191	.0195	0.9577	.4473	1.1691	14.0090**
4	COMMA	-.0181	.0133	1.8406	.4599	1.1641	11.2027**
5	DASH	.0329	.0250	1.7396	.4680	1.1621	9.3119**
6	MEANW	-.0174	.0121	2.0768	.4756	1.1602	8.0396**
7	SPELL	.0151	.0108	0.7457	.4828	1.1585	7.1233**
8	PROPN	.0133	.0129	1.0557	.4881	1.1582	6.3739**
9	RELPR	-.0247	.0194	1.6158	.4993	1.1573	5.7917**
10	PREP	-.0093	.0108	0.7457	.4967	1.1589	5.2736**
11	SDS	.0110	.0162	0.4583	.4988	1.1608	4.8198**

N=172.
Intercept: 2.9041

* Significant at .05 level
** Significant at .01 level

and five content-related. In this analysis series, the results of which are presented in Tables 4.33 through 4.36, the entire essay was used as the search unit for the five content variables.

When the content variables were included with the stylistic variables as predictors for the criterion measures on the science essays, the MULTR rose from .6010 (Table 4.29) to .6828 (Table 4.33). Length of essay and the proportion of spelling errors retained their significance as predictors; in addition, three of the five content variables ("teacher key phrases partially matched," "model student essay phrases completely matched," and "model student essay phrases partially matched") emerged as significant predictors.

Similarly, when the combination of stylistic and content variables was employed in a multiple regression analysis of the social studies essay group, the MULTR value achieved was .6809 -- compared with .6145 for the stylistic predictors alone. Length of essay was still the most effective single predictor of criterion scores for this group; however, the proportion of proper nouns was no longer significant in this group of predictors. As was true of the science essays, three of the five content variables have significant b-weights in this analysis; however, their late entrance into the multiple regression equa-

TABLE 4.33

STEPWISE MULTIPLE REGRESSION
SCIENCE ESSAYS
CONTENT AND STYLISTIC VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0367	.0080	20.9234**	.5382	.9919	90.9746**
2	KEYP	-.0228	.0097	5.4826*	.5972	.8698	61.5656**
3	MODC	-.0257	.0085	9.1692**	.6325	.8419	49.1303**
4	MODP	-.0200	.0100	3.9999*	.6485	.8252	39.9311**
5	COMMA	-.0109	.0062	3.0532	.6579	.8223	33.4354**
6	SPELL	.0140	.0059	5.5743*	.6648	.8175	28.7845**
7	SDW	-.0135	.0085	2.5242	.6734	.8109	25.7294**
8	PREP	.0065	.0063	1.0502	.6760	.8101	22.7313**
9	RELPR	.0064	.0060	1.1411	.6784	.8096	20.3772**
10	DASH	.0067	.0078	.7551	.6801	.8098	18.4205**
11	SDS	-.0058	.0066	.7661	.6815	.8102	16.8013**
12	KEYC	-.0046	.0080	.3341	.6820	.8116	15.3701**
13	MEANW	-.0033	.0077	.1875	.6824	.8131	14.1488**
14	OTHP	.0043	.0114	.1390	.6827	.8147	13.0959**
15	MEANS	.0014	.0069	.3987	.6828	.8166	12.1700**
16	PROPN	.0006	.0096	.0036	.6828	.8186	11.3552**

N=225.
Intercept: 3.2978

* Significant at .05 level
** Significant at .01 level

TABLE 4.34

STEPWISE MULTIPLE REGRESSION
SOCIAL STUDIES ESSAYS
CONTENT AND STYLISTIC VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0458	.0073	38.7347**	.5741	.8421	97.3546**
2	MEANW	.0030	.0093	.1077	.6035	.8222	56.4296**
3	SPELL	-.0008	.0059	.0188	.6414	.7931	45.6771**
4	RELPR	.0048	.0059	.6599	.6518	.7860	36.0181**
5	COMMA	-.0008	.0065	.0161	.6596	.7809	29.8942**
6	SDW	-.0081	.0087	.8645	.6662	.7769	25.6702**
7	PROPN	.0122	.0071	2.9847	.6697	.7757	22.3071**
8	PREP	.0080	.0065	1.5380	.6735	.7740	19.8295**
9	SDS	-.0025	.0065	.1592	.6764	.7733	17.8083**
10	DASH	.0107	.0065	3.7181	.6784	.7734	16.1160**
11	MEANS	.0046	.0066	.4890	.6795	.7744	14.6653**
12	KEYC	-.0213	.0088	5.8487*	.6803	.7757	13.4270**
13	KEYP	-.0123	.0088	1.9349	.6806	.7774	12.3506**
14	MODC	-.0064	.0098	.4258	.6808	.7793	11.4188**
15	MODP	-.0210	.0103	4.1459*	.6808	.7814	10.6018**
16	OTHP	.0178	.0081	4.8008*	.6809	.7835	9.8870**

N=200.
Intercept: 3.0987

* Significant at .05 level
** Significant at .01 level

tion indicates that these predictors, in the social studies essay group, account for very little variance among the essays by themselves, and are actually acting as suppressors because of their high correlation with the important length-of-essay variable.

The inclusion of the content variables in the multiple regression analysis of the English essays caused a rise in the MULTR from .5017 to .6225 (Table 4.35). In addition to the three significant stylistic predictors revealed in the earlier analysis (Table 4.31), three of the five content variables emerged as significant in this analysis -- the two variables reflecting counts of matches on the model student essays (MODC and MODP) and the count of matches on phrases from other sources (OTHP). A unique feature of the results here was that neither of the predictor variables associated with the teacher keys (KEYC and KEYP) proved to be significant predictors -- a situation not found among the other three subject areas. Another interesting result was that the length of essay variable, for these essays, was not so prominent a predictor as it was for the science and social studies essays (although it was a significant one).

Table 4.36 displays the results of the analysis of the for-

TABLE 4.35

STEPWISE MULTIPLE REGRESSION
ENGLISH ESSAYS
CONTENT AND STYLISTIC VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	MODC	-.0277	.0078	12.7634**	.3791	.8504	26.7018**
2	MODP	-.0212	.0084	6.4020*	.4632	.8355	21.5917**
3	FELPR	-.0289	.0076	14.6235**	.4935	.8088	16.8621**
4	PROPN	-.0133	.0077	3.0864*	.5301	.8237	15.2467**
5	LENGT	-.0102	.0076	5.7314*	.5592	.8079	14.1100**
6	MEANW	-.0119	.0089	1.7833	.5812	.7957	13.0948**
7	OTHP	.0377	.0185	4.1427*	.6006	.7848	12.3379**
8	KEYC	-.0187	.0104	3.2619	.6079	.7814	11.1410**
9	SDS	.0070	.0071	.9741	.6123	.7806	10.0666**
10	SPELL	.0060	.0070	.7326	.6157	.7806	9.1614**
11	COMMA	-.0062	.0070	.7822	.6181	.7814	8.3755**
12	SDW	-.0046	.0086	.2857	.6198	.7827	7.6936**
13	MEANS	.0045	.0073	.3748	.6207	.7846	7.0884**
14	KEYP	-.0055	.0096	.3360	.6218	.7864	6.5747**
15	DASH	-.0034	.0077	.1999	.6225	.7886	6.1168**
16	PREP	-.0005	.0071	.0043	.6225	.7913	5.6953**

N=161.
Intercept: 3.1755

* Significant at .05 level
** Significant at .01 level

TABLE 4.36

STEPWISE MULTIPLE REGRESSION
FOREIGN LANGUAGE ESSAYS
CONTENT AND STYLISTIC VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	KEYP	-.0747	.0114	42.7145**	.5107	1.1171	60.0128**
2	KEYC	-.0976	.0176	30.9198**	.6611	.9778	65.6228**
3	SDW	-.0267	.0103	6.7634*	.6762	.9629	47.1875**
4	COMMA	-.0217	.0110	3.8507	.6874	.9521	37.4050**
5	PROPN	.0161	.0109	2.1835	.6933	.9476	30.7381**
6	PELPR	-.0264	.0162	2.6586	.6971	.9456	26.0080**
7	SPELL	.0069	.0085	0.6645	.7003	.9443	22.5596**
8	MODP	-.0221	.0178	1.5397	.7030	.9437	19.9145**
9	PREP	-.0097	.0090	1.1717	.7050	.9440	17.7888**
10	LENGT	-.0229	.0167	1.8903	.7057	.9459	15.9783**
11	MEANS	.0195	.0167	1.3582	.7069	.9473	14.5327**
12	SDS	.0137	.0134	1.0442	.7085	.9482	13.3563**
13	DASH	.0154	.0208	0.5469	.7098	.9494	12.3563**
14	MEANW	.0020	.0104	0.0354	.7099	.9523	11.3940**
15	OTHP	.0080	.0458	0.0307	.7099	.9552	10.5708**

N=172.
Intercept: 2.9041

* Significant at .05 level
** Significant at .01 level

eign language essay group using both stylistic and content variables, with the essay as the search unit for determining the values of the content-related predictors. The multiple correlation coefficient obtained in this analysis was .7099. The two teacher key-related predictors were both highly effective here; standard deviation of word length was the only other significant variable. Recall that the MULTR for this group when the stylistic predictors only were employed was the lowest of the four subject areas (.4988). This subject area was also distinct from the other three in the lack of significance displayed by the length of essay variable.

The employment of the sentence as the search unit for content variable values in an analysis of the science essays (Table 4.37) produced a new significant variable for this subject area group -- the proportion of commas. Out of all the analyses done for the subject area groups, only in this one was the COMMA predictor significant. The final MULTR for this analysis was .6930, comparable to the .6828 value obtained for this essay group when the essay was used as the search unit.

When the sentence was used as the search unit for content variable matches in an analysis of the social studies essays (Table 4.38), only two predictors were significant -- length of

TABLE 4.37
STEPWISE MULTIPLE REGRESSION
SCIENCE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0405	.0078	27.2845**	.5382	.9119	90.9746**
2	NODP	-.0226	.0095	5.6558*	.6130	.8568	66.8273**
3	NODC	-.0244	.0079	9.5789**	.6477	.8281	53.2451**
4	COMMA	-.0123	.0061	4.0317*	.6580	.8203	42.0159**
5	KEYP	-.0199	.0094	4.4645*	.6656	.8148	34.8535**
6	SPELL	.0124	.0059	4.4597*	.6717	.8107	29.8773**
7	SDW	-.0124	.0084	2.1758	.6784	.8058	26.4394**
8	OTHP	.0218	.0131	2.7676	.6832	.8028	23.6431**
9	MEANS	.0064	.0069	.8595	.6863	.8015	21.2720**
10	DASH	.0089	.0077	1.3432	.6890	.8006	19.3411**
11	PREP	.0048	.0063	.5863	.6904	.8010	17.6368**
12	SDS	-.0047	.0065	.5208	.6912	.8020	16.1635**
13	RELPR	.0041	.0060	.4789	.6920	.8030	14.9187**
14	KEYC	-.0042	.0079	.2770	.6926	.8043	13.8306**
15	MEANW	-.0039	.0077	.2584	.6930	.8057	12.8795**
16	PROPN	.0008	.0009	.0065	.6930	.8077	12.0176**

N=225.
Intercept: 3.2978

* Significant at .05 level
** Significant at .01 level

TABLE 4.38
STEPWISE MULTIPLE REGRESSION
SOCIAL STUDIES ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step b-weight	SE	F-value	Cumulative Mult-R	SE	F-value
1	LENGT	-.0530	.0072	54.5753**	.5741	.8422	97.3547**
2	KEYC	-.0304	.0091	11.0299**	.6123	.8153	59.0875**
3	MODP	-.0204	.0106	3.7452	.6458	.7893	46.7630**
4	MEANS	.0088	.0068	1.6937	.6519	.7859	36.0363**
5	PROPN	.0111	.0073	2.2996	.6566	.7837	29.4110**
6	DASH	.0109	.0066	2.7264	.6597	.7829	24.7974**
7	SDW	-.0068	.0090	.5578	.6640	.7810	21.6381**
8	PREP	.0055	.0066	.7113	.6659	.7813	19.0279**
9	KEYP	-.0068	.0091	.5630	.6675	.7818	16.9723**
10	RELPR	.0043	.0060	.5050	.6687	.7828	15.2885**
11	CONJA	-.0001	.0057	.0416	.6688	.7848	13.8322**
12	OTHP	.0016	.0098	.0255	.6688	.7868	12.6158**
13	SPELL	.0005	.0061	.0057	.6688	.7889	11.5838**
14	MODC	-.0006	.0106	.0035	.6688	.7910	10.6990**
15	SDS	.0001	.0095	.0003	.6688	.7932	9.9318**
16	MEANW	.0000	.0095	.0000	.6688	.7953	9.2605**

N=200.
Intercept: 3.0988

* Significant at .05 level
** Significant at .01 level

essay and teacher key phrases completely matched. The final MULTR was .6686, compared with .6809 achieved when the essay search unit was used for this group of essays (Table 4.34).

In the analysis of the English essays in which the sentence was employed as the search unit for values on the content variables, four significant predictors appeared: model student essay phrases partially matched, model student essay phrases completely matched, length of essay, and proportion of relative pronouns (Table 4.39). The phrases from other sources partially matched (OTHP) and the proportion of proper nouns were not significant here as they had been in the corresponding analysis of the English essays in which the essay had been used as the search unit. The regression coefficient obtained in this analysis was .6241, compared with the .6225 MULTR achieved in the English essay analysis in which the essay search unit was employed (Table 4.35).

The results displayed in Table 4.40 (in which the sentence was used as the search unit for the content variables) and Table 4.36 (essay as search unit in an analysis of the foreign language area essays) are very similar. The same significant predictors appeared and the two final MULTR's are almost the same. These two MULTR's were the highest obtained in the entire study.

TABLE 4.39
STEPWISE MULTIPLE REGRESSION
ENGLISH ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	MODP	-.0226	.0078	8.4275**	.3531	.9603	22.6598**
2	MODC	-.0252	.0072	12.1858**	.4700	.8521	22.4000**
3	LENGT	-.0256	.0074	11.9337**	.5172	.8288	19.1210**
4	RELPR	-.0244	.0074	10.7540**	.5454	.8143	16.5142**
5	DMPN	-.0137	.0077	3.1518	.5765	.7953	15.4390**
6	MEANW	-.0079	.0090	.7663	.5915	.7884	13.8142**
7	QUAP	-.0273	.0193	2.0101	.6006	.7843	12.3376**
8	SDS	.0091	.0073	1.5634	.6088	.7807	11.1944**
9	COMMA	-.0080	.0070	1.3123	.6127	.7804	10.0844**
10	SPELL	.0060	.0070	.7489	.6159	.7805	9.1693**
11	MEANS	.0062	.0074	.7087	.6184	.7811	8.3906**
12	SDW	-.0055	.0086	.4171	.6203	.7823	7.7140**
13	KEYC	-.0066	.0087	.5754	.6219	.7837	7.1340**
14	DASH	-.0046	.0076	.3572	.6231	.7854	6.6189**
15	KEYP	-.0049	.0087	.3158	.6241	.7873	6.1669**
16	PROP	.0000	.0071	.0000	.6241	.7900	5.7484**

N=161.
Intercept: 3.1786

* Significant at .05 level
** Significant at .01 level

TABLE 4.[illegible]
STEPWISE MULTIPLE REGRESSION
FOREIGN LANGUAGE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	KEYP	-.0723	.0113	40.9478**	.5003	1.1251	56.7713**
2	KEYC	-.0999	.0167	35.6350**	.6568	.9826	64.1377**
3	SDW	-.0269	.0102	6.9982**	.6729	.9666	46.3606**
4	COMMA	-.0204	.0109	3.5002	.6837	.9556	36.6636**
5	PROPN	.0142	.0110	1.6710	.6895	.9524	30.0880**
6	MODP	-.0275	.0177	2.4111	.6937	.9500	25.5142**
7	OTHP	.0457	.0475	.9250	.6975	.9479	22.2077**
8	LENGT	-.0277	.0167	2.7560	.7000	.9476	19.5855**
9	SPELL	.0072	.0084	.7294	.7021	.9478	17.5022**
10	RELPR	-.0219	.0161	1.8496	.7048	.9471	15.8973**
11	PREP	-.0105	.0089	1.3998	.7066	.9473	14.5058**
12	MEANS	.0247	.0167	2.1715	.7090	.9475	13.3941**
13	SDS	.0150	.0132	1.2840	.7112	.9474	12.4446**
14	DASH	.0123	.0212	.3356	.7120	.9494	11.5347**
15	MEANW	.0009	.0102	.0078	.7120	.9524	10.6982**

N-172.
Intercept: 2.9041

* Significant at .05 level
** Significant at .01 level

Table 4.41 provides a summary of the twelve analyses of the essays by subject area. The MULTR's for the English and foreign language essays, which had the lowest prediction coefficients when only the eleven stylistic variables were used as predictors, made the greatest increase when the five content variables were included in the prediction equation. The highest MULTR values obtained in the study were found in this analysis series -- for the foreign language area essays when both content and stylistic variables were employed in the prediction equation. In this analysis set, the poorest overall performance of these predictors occurred when the English area essays were analyzed. The resulting MULTR values for all subject areas when both content and stylistic variables were used were quite similar whether the essay or the sentence was chosen as the search unit for values for the content variables. In all eight cases where a comparison was made between the "stylistic only" MULTR for a subject area and the MULTR's (essay or sentence as search unit) in which both content and stylistic variables were included in the prediction equation, the $F_{difference}$ values between the two coefficients were determined to be highly significant.

TABLE 4.41
SUMMARY OF ANALYSES BY SUBJECT AREA

Subject Area		Mult-R's Stylistic Variables Only	Mult-R's Stylistic and Content Variables (Essay as Unit)	F_1	Mult-R's Stylistic and Content Variables (Sentence as Unit)	F_2
Science	Original	.6010	.6828		.6930	
N=225	Shrunken	(1).5756**	(2).6541**	7.1481**	(3).6655**	8.2593**
	Attenuated	.6307	.7168	(1) vs. (2)	.7293	(1) vs. (3)
Social Studies						
N=208	Original	.6145	.6809		.6688	
	Shrunken	(4).5871**	(5).6479**	4.8387**	(6).6343**	3.4848**
	Attenuated	.6433	.7100	(4) vs. (5)	.6951	(4) vs. (6)
English	Original	.5017	.6225		.6241	
N=161	Shrunken	(7).4493**	(8).5693**	5.3043**	(9).5713**	5.2979**
	Attenuated	.4923	.6238	(7) vs. (8)	.6261	(7) vs. (9)
Foreign Language						
N=172	Original	.4988	.7099		.7120	
	Shrunken	(10).4496**	(11).6754**	14.5142**	(12).6779**	14.7143**
	Attenuated	.4927	.7401	(10) vs. (11)	.7429	(10) vs. (12)

** Significant at .01 level

Analyses by Grade Level

Three grade level groups had been identified for the sample of essays under study -- 7th-8th, 9th-10th, and 11th-12th. The number of essays available for the three levels was 286, 245, and 227, respectively. Three series of analyses, each employing a different combination of predictor variables, were carried out for each of these three groups. The variable sets were:

- stylistic variables alone (eleven predictors)
- stylistic and content variables; essay as search unit for content variable values (sixteen predictors)
- stylistic and content variables; sentence as search unit for content variable values (sixteen predictors)

For the 7th-8th grade level group, only length of essay proved to be a significant predictor in the multiple regression analysis using stylistic variables only (Table 4.42). A final MULTR of .4932 was obtained in this analysis.

At the next grade level (9th-10th), the proportion of spelling errors appeared along with essay length as a significant predictor, and the MULTR reached a value of .5787 (Table 4.43).

As shown in Table 4.44, a MULTR value of .4882 was achieved for the 11th-12th grade level essay group in the analysis using stylistic variables only. As in the analyses for the other two

TABLE 4.42
STEPWISE MULTIPLE REGRESSION
SEVENTH AND EIGHTH GRADE ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0349	.0084	42.6653**	.4441	1.0469	69.8077**
2	COMMA	-.0138	.0071	3.7328	.4691	1.0338	39.9417**
3	SDW	-.0125	.0084	2.2574	.4823	1.0273	28.5007**
4	SPELL	.0068	.0065	1.0961	.4862	1.0265	21.7517**
5	DASH	.0073	.0083	.7369	.4877	1.0274	17.4831**
6	SDS	.0056	.0077	.5242	.4893	1.0282	14.6381**
7	PREP	-.0057	.0070	.6498	.4905	1.0292	12.5876**
8	MEANW	-.0076	.0089	.7229	.4926	1.0297	11.0980**
9	RELPR	.0028	.0075	.1432	.4931	1.0312	9.8541**
10	MEANS	-.0012	.0075	.0246	.4932	1.0330	8.8401**
11	PROPN	-.0012	.0082	.0200	.4932	1.0349	8.0097**

N=288.
Intercept: 3.2657

** Significant at .01 level

TABLE 4.43
STEPWISE MULTIPLE REGRESSION
NINTH AND TENTH GRADE ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0553	.0062	79.0879**	.5238	.8813	91.8830**
2	MEANW	-.0118	.0076	2.4190	.5434	.8703	50.7168**
3	SPELL	.0164	.0059	7.6488**	.5635	.8582	37.3771**
4	DASH	-.0039	.0078	.2484	.5673	.8573	28.4806**
5	PROPN	.0085	.0072	1.3809	.5706	.8567	23.0855**
6	PREP	.0073	.0062	1.3912	.5733	.8565	19.4259**
7	SDW	-.0085	.0076	1.2497	.5759	.8564	16.8097**
8	SDS	.0065	.0067	.9334	.5786	.8562	14.8484**
9	COMMA	.0007	.0066	.0111	.5786	.8580	13.1458**
10	MEANS	.0009	.0007	.0157	.5787	.8598	11.7826**
11	RELPR	-.0005	.0067	.0061	.5787	.8617	10.6665**

N=245.
Intercept: 3.0286

** Significant at .01 level

TABLE 4.44

STEPWISE MULTIPLE REGRESSION
ELEVENTH AND TWELFTH GRADE ESSAYS
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0431	.0079	29.3725**	.4097	.9880	45.4001**
2	MEANW	-.0151	.0100	2.2801	.4503	.9692	28.4870**
3	DASH	.0171	.0085	4.0873	.4644	.9635	20.4439**
4	SPELL	.0112	.0071	2.4808	.4726	.9608	15.9805**
5	SDW	-.0130	.0106	1.4472	.4778	.9600	13.0807**
6	RELPR	-.0076	.0075	1.0243	.4824	.9594	11.1245**
7	CONJA	-.0077	.0078	.9844	.4866	.9590	9.7103**
8	PROPN	.0055	.0090	.3805	.4879	.9605	8.5135**
9	SDS	-.0021	.0075	.0793	.4882	.9625	7.5457**
10	MEANS	.0009	.0080	.0139	.4882	.9647	6.7615**
11	PREP	-.0002	.0074	.0011	.4882	.9669	6.1185**

N=227.
Intercept: 3.0683

* Significant at .05 level
** Significant at .01 level

levels, the length of essay was by far the most effective predictor variable, accounting by itself for approximately 17% of the criterion variance. The proportion of dashes (DASH) predictor was also significant for this group; this was the only analysis done with the essays in which the DASH predictor was significant.

The addition of the content variables to the independent variable group in an analysis of the 7th-8th grade level essays increased the MULTR from .4932 to .6144 (Table 4.45). This significant increase was achieved largely due to the effective prediction of three of the new variables -- teacher key phrases partially matched, model student essay phrases completely matched, and other source pheases partially matched. For this analysis of the lowest grade level group, the essay as a whole was used as the search unit for determining values for the content variables. The remaining thirteen predictors raised the MULTR only about .02 over the level attained by the first four variables entered.

Table 4.46 reveals that the content variables related to teacher key phrases (KEYC and KEYP) proved significant for the 9th-10th grade level essays when the five content variables (values formed by searching through the essay as a whole) were included in the multiple regression equation. The two significant stylistic variables from the previous analysis for this

TABLE 4.43

STEPWISE MULTIPLE REGRESSION
SEVENTH AND EIGHTH GRADE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	KEYP	-.0376	.0085	19.7561**	.4610	1.0369	76.6721**
2	LENGT	-.0316	.0083	14.6737**	.5400	.9852	58.2657**
3	MODC	-.0248	.0098	6.3921*	.5698	.9636	45.2131**
4	OTHP	.0347	.0147	6.5046*	.5896	.9488	37.4444**
5	CONNA	-.0127	.0067	3.6323	.5999	.9415	31.4899**
6	NODP	-.0182	.0097	3.5353	.6071	.9368	27.1541**
7	SEW	-.0105	.0078	1.8277	.6099	.9359	23.5332**
8	DASH	.0079	.0077	1.0565	.6119	.9358	20.7291**
9	SDS	.0045	.0071	.3942	.6129	.9366	18.4546**
10	KEYC	-.0055	.0093	.3409	.6137	.9376	16.6167**
11	PREP	-.0025	.0065	.1462	.6140	.9389	15.0807**
12	PROPN	-.0022	.0076	.0826	.6142	.9405	13.7825**
13	RELPR	-.0016	.0069	.0560	.6143	.9422	12.6831**
14	MEANS	.0019	.0083	.0526	.6144	.9438	11.7404**
15	SPELL	.0001	.0060	.0006	.6144	.9455	10.9173**
16	MEANS	-.0001	.0072	.0001	.6144	.9473	10.1971**

N=266.
Intercept: 3.2657

* Significant at .05 level
** Significant at .01 level

TABLE 5.46
STEPWISE MULTIPLE REGRESSION
NINTH AND TENTH GRADE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0354	.0067	28.2084**	.5238	.8813	91.8830**
2	KEYP	-.0212	.0081	6.8409**	.5586	.8599	54.8850**
3	KEYC	-.0195	.0078	6.2746*	.6006	.8260	46.7574**
4	SPELL	.0130	.0056	6.0855*	.6193	.8181	37.1463**
5	SDW	-.0096	.0071	1.8317	.6313	.8091	31.6740**
6	PREP	.0062	.0058	1.1527	.6571	.8957	27.1084**
7	NODC	-.0154	.0090	2.9211	.6419	.8033	23.7298**
8	NODP	-.0142	.0084	2.8222	.6494	.7983	21.5139**
9	PROPN	.0115	.0069	2.8789	.6631	.7966	19.4312**
10	OTHP	.0117	.0136	.7427	.6551	.7965	17.6011**
11	MEANW	-.0059	.0071	.6823	.6568	.7966	16.0765**
12	SDS	.0043	.0064	.4660	.6578	.7975	14.7504**
13	DASH	-.0026	.0075	.1218	.6581	.7990	13.5753**
14	RELPR	-.0020	.0063	.1025	.6582	.8005	12.5641**
15	COMMA	-.0014	.0062	.0491	.6583	.8022	11.6810**
16	MEANS	.0004	.0065	.0033	.6583	.8039	10.9034**

N=245
Intercept: 3.0286

* Significant at .05 level
** Significant at .01 level

grade level (LENGT and SPELL) remained so in this analysis, and the MULTR rose from .5787 to .6583.

As in the previous analysis, the two teacher key-related content variables emerged as significant predictors in the analysis of 11th-12th grade essays using both stylistic and content variables, with the essay as search unit for the content variable values (Table 4.47). A MULTR of .6233 was obtained, a rise of .04 over the value reached when the stylistic variables alone were used. The length of essay remained highly significant, but with this combination of predictors, the proportion of dashes no longer contributed significantly to the value of the MULTR, as it had in the stylistic analysis reported in Table 4.44.

Table 4.48 presents the results of an analysis of the 7th-8th grade level essays, again using all sixteen predictor variables (as in the analysis reported in Table 4.45), but here with each student essay sentence used as the search unit for building values for the five content variables. The achieved MULTR here is slightly lower than for the comparable analysis in which the entire essay was used as the search unit for the content variables (Table 4.45) -- .5952 vs. .6144. The proportion of commas, which had not been a significant predictor in the previous analyses of this grade level group, was signifi-

TABLE 4.07

STEPWISE MULTIPLE REGRESSION
ELEVENTH AND TWELFTH GRADE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0223	.0080	7.8519**	.4097	.9870	45.4001**
2	KEYC	-.0556	.0103	28.8590**	.4634	.9619	30.6363**
3	KEYP	-.0382	.0090	18.0686**	.5706	.8934	35.8916**
4	MEANV	-.0129	.0093	1.9411	.5852	.8841	28.9112**
5	DASH	.0148	.0076	3.6137	.5963	.8773	24.3916**
6	MODP	-.0167	.0098	2.9071	.6031	.8736	20.9678**
7	RELFR	-.0115	.0068	2.8718	.6086	.8710	18.4120**
8	MEANS	.0097	.0075	1.6610	.6133	.8691	16.4335**
9	COMMA	-.0094	.0071	1.7624	.6179	.8671	14.8985**
10	OTHP	.0087	.0092	.9041	.6201	.8672	13.4973**
11	STS	-.0046	.0068	.4384	.6214	.8680	12.2987**
12	SDW	-.0048	.0101	.2254	.6220	.8695	11.2565**
13	SPELL	.0032	.0066	.2349	.6226	.8711	10.3725**
14	PREP	-.0026	.0067	.1538	.6230	.8728	9.6060**
15	MODC	-.0037	.0094	.1557	.6233	.8746	8.9393**
16	PROPN	-.0010	.0083	.0147	.6233	.8766	8.3423**

N=227
Intercept: 3.0683

** Significant at .01 level

TABLE 4.48
STEPWISE MULTIPLE REGRESSION
SEVENTH AND EIGHTH GRADE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1*	LENGT	-.0362	.0084	18.5203**	.4441	1.0469	69.8077**
2	KEYP	-.0341	.0087	15.2360**	.5257	.9957	54.0638**
3	VOCC	-.0255	.0103	6.2053*	.5569	.9720	42.6560**
4	COMMA	-.0135	.0067	4.0589*	.5711	.9643	37.0167**
5	VOCP	-.0180	.0098	3.3745	.5787	.9798	28.2027**
6	SDV	-.0136	.0080	2.9057	.5841	.9569	24.0835**
7	DASH	.0090	.0078	1.3374	.5870	.9562	20.8816**
8	KEYC	-.0114	.0093	1.4552	.5899	.9556	18.4587**
9	COMP	-.0143	.0145	.9751	.5911	.9560	16.4759**
10	PREP	-.0048	.0066	.5213	.5922	.9568	14.8596**
11	SDS	.0052	.0073	.5070	.5931	.9578	13.5194**
12	MEANS	.0053	.0073	.5216	.5941	.9587	12.4106**
13	HELPR	-.0035	.0071	.2478	.5946	.9600	11.4462**
14	MEANI	.0032	.0085	.1432	.5949	.9615	10.6083**
15	SKILL	.0019	.0061	.0975	.5952	.9630	9.8763**
16	PROPN	-.0012	.0078	.0235	.5952	.9648	9.2270**

N = 96
Intercept: 3.2687

* Significant at .05 level
** Significant at .01 level

cant for this analysis, while the count of partial matches on the "other source" phrases (OTHP), which had been significant in the analysis where the essay unit was employed, was not significant in the results reported in Table 4.48 -- indicating that the contingency restraints placed on the matching possibilities by the smaller search unit produced counts on this variable which did not correlated with the criterion measure, perhaps either by making the OTHP values more uniform across essays or by preventing many matches on the variable.

The new search unit (essay sentence) also produced a change in the significant predictors for the 9th-10th grade level essays (Table 4.49). The prominence of the teacher key-related predictors was reduced, while both the model student essay-related variables were now significant (neither of them had been in the corresponding analysis reported in Table 4.46). Length of essay and the proportion of spelling errors were significant in this analysis, as they had been for each of the other analyses for this grade level group. The MULTR attained when the boundaries of the essay sentence determined counts on the content variables was .6792, compared to .6583 achieved with the essay as search unit.

A new significant stylistic variable, the proportion of relative pronouns (RELPR), appeared among the predictors for the 11th-

TABLE 5.40
STEPWISE MULTIPLE REGRESSION
NINTH AND TENTH GRADE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENST	-.0394	.0052	40.6165**	.5238	.4813	91.8830**
2	MODP	-.0247	.0075	10.7458**	.6032	.8269	69.2133**
3	MODC	-.0170	.0074	5.2603*	.6373	.7954	55.0070**
4	SIMOL	.0132	.0055	5.8202*	.6493	.7858	43.1647**
5	SAC	-.0116	.0070	2.7075	.6560	.7803	35.9235**
6	MOVP	-.0161	.0076	4.3139*	.6602	.7852	30.6521**
7	PROPN	.0118	.0066	3.2412	.6642	.7831	26.7318**
8	CZIP	.0309	.0200	2.3737	.6650	.7812	23.7730**
9	MOVC	-.0092	.0066	1.9785	.6710	.7794	21.4647**
10	SOS	.0063	.0061	1.0691	.6730	.7779	19.5879**
11	PREP	.0048	.0057	.7123	.6776	.7770	17.9864**
12	MEANS	.0044	.0063	.4903	.6781	.7782	16.4625**
13	RELPR	-.0039	.0062	.3956	.6788	.7791	15.1899**
14	MEANN	-.0024	.0071	.1107	.6791	.7806	14.0627**
15	CONJA	-.0017	.0061	.0753	.6792	.7822	13.0771**
16	DASH	-.0011	.0073	.0231	.6792	.7838	12.2089**

N=245
Intercept: 3.0286

* Significant at .05 level
** Significant at .01 level

12th grade level essays when an analysis was done using student essay sentences as the search units for matches on the essay key phrases (Table 4.50). The other three significant variables in this analysis -- length of essay, teacher key phrases partially matched, and teacher key phrases completely matched -- were those which had been significant in the corresponding analysis (Table 4.47) using all sixteen predictors for this grade level group. The MULTR value for this analysis was .6319, a slight increase over the corresponding analysis reported in Table 4.47.

Following the three sets of analyses of the essays by grade level, shrunken MULTR's and MULTR values corrected for attenuation for each of the nine original MULTR values were estimated mathematically. These values are presented in Table 4.51. As the table indicates, the inclusion of the five content variables in the multiple regression equations significantly increased the multiple correlation coefficients in all cases. The F_1 values reported are measures of the significance of the difference between the MULTR values obtained when the full complement of available stylistic and content variables were used as predictors (with the entire essay employed as the search unit for the content variable values), and the MULTR values obtained when

TABLE 4.50
STEPWISE MULTIPLE REGRESSION
ELEVENTH AND TWELFTH GRADE ESSAYS
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0258	.0077	11.3664**	.4097	.9880	45.4001**
2	MEYC	-.0599	.0102	33.6560**	.4677	.9594	31.3658**
3	MEMT	-.0307	.0087	20.6781**	.5780	.8878	37.2997**
4	MEWM	-.0147	.0092	2.5684	.5916	.8791	29.8852**
5	DASH	.0120	.0077	2.4190	.6016	.8729	25.0724**
6	RELPR	-.0133	.0068	3.8505*	.6093	.8685	21.6579**
7	MODP	-.0138	.0094	2.1627	.6157	.8650	19.1085**
8	MEANS	.0117	.0074	2.5185	.6218	.8617	17.1828**
9	COMMA	-.0115	.0071	2.5991	.6275	.8587	15.6646**
10	OTHP	.0104	.0101	1.0468	.6296	.8588	14.1909**
11	SPS	-.0048	.0068	.4946	.6309	.8596	12.9297**
12	SPELL	.0035	.0065	.2955	.6316	.8610	11.8397**
13	PAEP	-.0021	.0067	.0973	.6319	.8628	10.8915**
14	SYN	-.0015	.0100	.0240	.6319	.8647	10.0692**
15	PROPN	.0007	.0082	.0032	.6319	.8668	9.3544**
16	NODC	.0005	.0091	.0025	.6319	.8688	8.7285**

N=227
Intercept: 3.0683

* Significant at .05 level
** Significant at .01 level

TABLE 4.51
SUMMARY OF ANALYSES BY GRADE LEVEL

Grade Level		Mult-R's Stylistic Variables Only	Mult-R's Stylistic and Content Variables (Essay as Unit)	F_1	Mult-R's Stylistic and Content Variables (Sentences as Unit)	F_2
7th-8th	Original	.4932	.6144		.5952	
N=286	Shrunken	(1) .4644**	(2) .5856**	10.5833**	(3) .5643**	8.2000**
	Attenuated	.5089	.6417	(1) vs. (2)	.6184	(1) vs. (3)
9th-10th	Original	.5787	.6583		.6792	
N=245	Shrunken	(4) .5536**	(5) .6295**	6.8846**	(6) .6526**	9.5200**
	Attenuated	.6066	.6898	(4) vs. (5)	.7151	(4) vs. (6)
11th-12th	Original	.4882	.6233		.6319	
N=227	Shrunken	(7) .4506**	(8) .5874**	9.1612**	(9) .5972**	10.2333**
	Attenuated	.4938	.6437	(7) vs. (8)	.6544	(7) vs. (9)

** Significant at .01 level

only stylistic variables were used. F_2 values indicate the significance of the difference between the MULTR values for the analyses using the combination of stylistic and content variables with the sentence used as the search unit for matches on the content variables and the "stylistic only" MULTR's. All $F_{difference}$ values were highly significant.

No trend in the size of the regression coefficients can be detected as the grade level rises. The lowest MULTR value in the stylistic analyses was found for the 11th-12 grade level essays, evidence of greater rater emphasis on essay content at the highest grade level. However, the highest stylistic multiple correlation coefficient is associated with the 9th-10th grade level group, the same group for which the best prediction was achieved when the entire group of predictors was used. The lowest overall performance of the predictors occurred with the 7th-8th grade level essays. Here the MULTR value when stylistic variables only were employed was almost as low as the lowest stylistic MULTR -- that for the 11th-12th grade group; also, for the 7th-8th grade essays, the lowest MULTR's for the combination of variables were obtained, both when the entire essay was used as the search unit and when the sentence was used as the search unit

for scores representing matches on the content variables.

Only in the case of the 9th-10th grade level did the shrunken MULTR's for the combination of variables exceed .60; however, the attenuated values for these MULTR's were over .60 in all cases.

Analyses by Content Specificity Level

Five raters for each of the four subject areas had been asked to place each of the twelve essay questions available for each subject area along a ten-point continuum, with one end of the continuum representing "high content specificity" (i.e., content requiring emphasis in evaluating essay response) and the other end of the continuum representing low content specificity (stylistic considerations requiring emphasis in evaluating essay response). The procedures used for this rating were outlined in Chapter III; a copy of the instructions given to raters is presented in Appendix E.

Using the average ratings obtained for each essay item, the 48 questions were rank-ordered from highest mean rating to lowest and then divided evenly into three groups representing high, middle, and low content specificity. Table 4.52 shows the grouping of essay questions on the basis of content specificity, indicating for each item the subject area represented by it, and its rank based on the mean content specificity rating given to the item. Note that the majority of foreign language essays (subject area code #4) fall into the high content specificity group, with no foreign language items in the low group. In contrast, most English essay items are contained in the low content specificity

TABLE 4.52
GROUPING OF ESSAY QUESTIONS BY CONTENT SPECIFICITY

High			Middle			Low		
Quest. #	Subj. Area	Rank	Quest. #	Subj. Area	Rank	Quest. #	Subj. Area	Rank
1421	4	1	1221	2	17	1321	3	33
2421	4	1	2211	2	17	2321	3	33
1422	4	3	3111	1	19	2322	3	35
2422	4	3	1411	4	20	3331	3	37
1111	1	5	3112	1	21	2112	1	37
1112	1	6	3421	4	22	1211	2	37
2411	4	7	2312	3	22	2311	3	39
1121	1	9	2121	1	24	2111	1	40
3121	1	9	3212	2	25	3222	2	41
3122	1	9	3211	2	25	1311	3	42
2221	2	11	1122	1	27	1322	3	42
3422	4	12	2212	2	28	1212	2	44
1412	4	13	2122	1	30	3221	2	45
2412	4	13	3411	4	30	3332	3	46
1222	2	15	3412	4	30	3321	3	47
2222	2	15	1312	3	32	3322	3	47

Subject Area Code

1 - Science
2 - Social Studies
3 - English
4 - Foreign Language

group; none were placed in the high group. Social studies items are listed only under the middle and low content specificity categories, while science items are represented in each of the three categories.

A number of analyses, using essay responses for all items under each category to compose three comparison groups, were carried out; the results of these analyses are shown in Tables 4.53 through 4.64.

Four significant predictors were apparent in the stylistic analysis of the high content specificity group (Table 4.53). These were length of essay, mean word length, proportion of commas, and standard deviation of word length. The MULTR value attained was .5161.

For the "middle" content specificity level group, a MULTR of .6887 was achieved (Table 4.54), the highest value for a stylistic analysis of any group. Only two variables were significant in this analysis -- length of essay and the proportion of spelling errors, with length of essay by far the most effective single predictor.

The group of essays whose content was judged least specific produced the lowest stylistic MULTR of the three groups (Table 4.55) -- an unexpected result, since it was thought that those essay questions for which response content was most flexible

TABLE 4.53
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- HIGH GROUP
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0475	.0103	21.1139**	.4090	1.1783	59.2796**
2	MEANW	-.0256	.0091	7.9548**	.4727	1.1398	42.3155**
3	COMMA	-.0187	.0088	4.5229*	.4893	1.1300	30.7465**
4	SDW	-.0178	.0090	3.9561*	.4985	1.1251	24.1498**
5	SPELL	.0124	.0069	3.2106	.5063	1.1212	20.0635**
6	PROPN	.0102	.0097	1.1088	.5096	1.1205	16.9620**
7	DASH	.0127	.0108	1.3859	.5132	1.1197	14.7646**
8	MEANS	-.0072	.0097	.5468	.5153	1.1200	13.0154**
9	PREP	-.0044	.0073	.3732	.5163	1.1212	11.5902**
10	RELPR	.0014	.0085	.0268	.5163	1.1231	10.3986**
11	SDS	-.0009	.0093	.0088	.5163	1.1251	9.4213**

N=287.
Intercept: 3.1650

* Significant at .05 level
** Significant at .01 level

TABLE 4.54
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- MIDDLE GROUP
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0592	.0061	92.8135**	.5613	.8125	120.1075**
2	SPELL	.0109	.0053	4.2819*	.5729	.8093	63.5210**
3	COMMA	-.0087	.0055	2.5234	.5820	.8014	44.2446**
4	RTLTR	.0075	.0056	1.7842	.5885	.7984	34.1837**
5	SLN	-.0122	.0077	2.5491	.5933	.7964	27.9296**
6	PREP	.0072	.0055	1.6829	.5963	.7958	23.5490**
7	SDS	.0069	.0058	1.4194	.5989	.7955	20.3744**
8	MEANW	.0063	.0074	.7280	.6004	.7959	17.9012**
9	PROPN	.0019	.0062	.0927	.6006	.7973	15.8680**
10	ENSE	.0021	.0067	.0932	.6008	.7987	14.2371**
11	MEANS	.0015	.0060	.0605	.6009	.8002	12.9000**

N=263.
Intercept: 2.8973

* Significant at .05 level
** Significant at .01 level

TABLE 4.55
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- LOW GROUP
STYLISTIC VARIABLES ONLY

Step	Variable Entered	Final Step b-weight	SE	F-value	Cumulative Mult-R	SE	F-value
1	LENGT	-.0421	.0062	45.9344**	.4080	.9031	46.3958**
2	MEANN	-.0157	.0085	3.3635	.4353	.8929	26.8913**
3	POLIR	-.0148	.0064	5.3986*	.4579	.8838	20.2213**
4	MEAN3	.0101	.0068	2.2076	.4725	.8779	16.3803**
5	SPELL	.0077	.0061	1.5946	.4788	.8775	13.3655**
6	COMMA	.0063	.0035	.9414	.4801	.8777	11.2885**
7	DASH	.0067	.0068	.9966	.4833	.8779	9.7982**
8	PROPN	.0051	.0066	.5895	.4854	.8787	8.6339**
9	PREP	.0045	.0056	.4663	.4869	.8798	7.6994**
10	SPL	-.0041	.0083	.2563	.4877	.8813	6.9310**
11	SDS	.0004	.0068	.0031	.4877	.8833	6.2729**

N=232.
Intercept: 3.1738

* Significant at .05 level
** Significant at .01 level

would also be the ones for which the raters would emphasize the stylistic features of the responses in their judgment of essay quality. Two predictors were significant -- length of essay and the proportion of relative pronouns. The final MULTR value was .4877. Recall that four English essay questions had been deleted from all the analyses due to lack of any specific content required by the questions; no computer-usable key could be composed for such items, since there was no one set of words or phrases which would cover the information required in the student responses to these questions. Obviously, all four of these English questions had been placed in the lowest content specificity category by the raters. Therefore, the group of essays comprising the low content specificity group consisted only of responses to eight out of the twelve essay questions which had originally been placed in that category.

The two teacher key-related content variables were the most significant predictors for the high content specificity group when both stylistic and content variables, with the essay as search unit, were employed in the multiple regression analysis (Table 4.56). Length of essay and standard deviation of word length remained significant, as they had been in the stylistic analysis of the group (Table 4.53); however, mean word length and proportion of commas were no longer significant. The final

TABLE 4.56
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- HIGH GROUP
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	KEYP	-.0616	.0100	37.7671**	.4833	1.1305	89.9280**
2	KEYC	-.0362	.0103	13.8885**	.5991	1.0356	82.3421**
3	LENGT	-.0218	.0100	4.7662*	.6162	1.0205	59.7899**
4	SDC	-.0191	.0082	5.4749*	.6297	1.0082	47.9878**
5	COMMA	-.0109	.0081	1.8370	.6350	1.0043	39.3345**
6	MODC	-.0171	.0130	1.7443	.6379	1.0029	33.1783**
7	SPELL	.0085	.0063	1.8210	.6406	1.0018	28.7414**
8	PROPN	.0119	.0089	1.8001	.6431	1.0074	25.3985**
9	DASH	.0117	.0098	1.4326	.6453	1.0001	22.7603**
10	RELPR	-.0062	.0077	.6450	.6465	1.0005	20.5448**
11	MEANW	-.0052	.0085	.3719	.6473	1.0013	18.6931**
12	SDS	-.0020	.0084	.0534	.6475	1.0029	17.0907**
13	PREP	.0016	.0069	.0561	.8476	1.0045	15.7297**
14	QUOP	-.0026	.0124	.0435	.6477	1.0062	14.5591**
15	MEANS	.0009	.0090	.0109	.6477	1.0080	13.5417**
16	MODP	-.0010	.0120	.0069	.6477	1.0098	12.6509**

N=287.
Intercept: 3.1650

* Significant at .05 level
** Significant at .01 level

MULTR for this analysis was .6177, compared to .5163 for the analysis in which only the eleven stylistic variables had been used. This rise of .13 in the multiple regression coefficient was the greatest increase found among the three content specificity groups between the "stylistic only" analysis and the analysis in which the full complement of sixteen predictors was employed. Note that the effective predictors here were the same as the ones obtained when an analysis using both content and stylistic variables was done for the foreign language essay group (Table 4.36) with one exception: in the analysis of the foreign language essays, the length of essay variable was not significant; also, a higher prediction level (.7099 with the essay as search unit) was achieved in that analysis.

Table 4.57 shows the results obtained when all sixteen independent variables (essay as search unit for the content variable values) were used in an analysis of the middle content specificity group. A multiple correlation coefficient of .6565 was obtained and three significant predictors were apparent: the ubiquitous length of essay, teacher key phrases completely matched, and model student essay phrases partially matched. The MULTR for this analysis showed the least rise over the corresponding stylistic analysis for any of the three groups.

TABLE 4.57
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- MIDDLE GROUP
STYLISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	LENGT	-.0455	.0066	47.6893**	.5613	.8125	120.1075**
2	KEYC	-.0192	.0078	6.1108*	.5867	.7966	68.2592**
3	MODP	-.0212	.0080	7.0054**	.6206	.7728	54.0907**
4	RELPR	.0066	.0054	1.4973	.6294	.7674	42.3237**
5	COMMA	-.0095	.0053	3.1848	.6350	.7643	34.7386**
6	KEYP	-.0110	.0072	2.3324	.6411	.7608	29.7789**
7	SPELL	.0075	.0051	2.2069	.6462	.7580	26.1229**
8	MODC	-.0116	.0084	1.8884	.6497	.7566	23.1954**
9	PREP	.0049	.0053	.8324	.6515	.7565	20.7405**
10	SLN	-.0087	.0075	1.3669	.6526	.7571	18.6966**
11	OTHP	.0089	.0116	.5862	.6537	.7577	17.0279**
12	MEANW	.0053	.0071	.5496	.6544	.7586	15.6050**
13	MEANS	.0044	.0058	.5675	.6550	.7595	14.3974**
14	SEG	.0038	.0056	.4548	.6558	.7604	13.3684**
15	DASH	.0039	.0065	.3694	.6564	.7613	12.4713**
16	PROPN	.0010	.0060	.0272	.6565	.7628	11.6475**

N=263.
Intercept: 2.8073

* Significant at .05 level
** Significant at .01 level

Using the essay as search unit for matches on the content variables, an analysis was performed for the low content specificity group with all sixteen predictors allowed to enter the equation. The results of this analysis are shown in Table 4.58. Neither of the teacher key-related variables were significant; however, the other three content variables were significant. Length of essay and proportion of relative pronouns were significant, as they had been in the analysis of this group in which stylistic variables only were employed (Table 4.55). The resulting MULTR in this analysis was .6002, over .4877 for the corresponding stylistic analysis.

Contrasting Table 4.59 with Table 4.56 discloses that an additional variable, "model student essay phrases completely matched," was significant only when the sentence, rather than the essay as a whole, was used as the search unit for content variable values. The two resulting MULTR's were almost identical.

As in the analysis described immediately above, the presence of the MODC variable as a significant predictor distinguishes the two analyses for the middle content specificity group in which both stylistic and content variables were used (Table 4.60), this variable appearing as a significant predictor only when the sentence was used as the search unit for matches on the content variables. The MULTR obtained for the analysis of the middle content speci-

TABLE 4.58
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- LOW GROUP
STATISTIC AND CONTENT VARIABLES
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Multi-R	SE	F-value
1	LENGT	-.0274	.0064	18.6425**	.4089	.9031	46.3958**
2	MODC	-.0231	.0071	10.5552**	.4784	.8710	34.1341**
3	MODP	-.0214	.0078	7.5588**	.5321	.8516	30.1597**
4	CTHP	.0083	.0101	10.6273**	.5577	.8268	25.7432**
5	RELPR	-.0171	.0060	8.2100**	.5775	.8150	22.7241**
6	MEANW	-.0124	.0081	2.3658	.5875	.8097	19.8571**
7	KEYC	-.0131	.0083	2.5004	.5923	.8079	17.3749**
8	CONJA	.0067	.0061	1.2345	.5956	.8073	15.3942**
9	MEANS	.0046	.0065	.5069	.5973	.8079	13.7445**
10	KEYP	-.0057	.0085	.4418	.5984	.8088	12.3863**
11	DASH	.0033	.0063	.2784	.5991	.8101	11.2494**
12	PREP	.0030	.0061	.2352	.5996	.8115	10.2938**
13	SPELL	.0022	.0058	.1459	.6000	.8131	9.4772**
14	PROPN	.0017	.0062	.0717	.6002	.8149	8.7684**
15	SDS	.0006	.0063	.0083	.6002	.8167	8.1471**
16	SDW	-.0001	.0077	.0001	.6002	.8186	7.6027**

N=233.
Intercept: 3.1738

** Significant at .01 level

TABLE 4.59
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- HIGH GROUP
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
1	KEYP	-.0570	.0010	32.3769**	.4721	1.1383	84.6118**
2	KEYC	-.0424	.0104	16.5147**	.5866	1.0476	77.1379**
3	LENGT	-.0291	.0097	9.0699**	.6085	1.0281	57.4541**
4	SLW	-.0203	.0082	6.0989*	.6234	1.0148	46.4202**
5	NODC	-.0284	.0139	4.1669*	.6306	1.0090	38.4316**
6	SPELL	.0087	.0063	1.9265	.6339	1.0073	32.4703**
7	COMMA	-.0093	.0081	1.3296	.6363	1.0064	28.0932**
8	DASH	.0133	.0097	1.8589	.6385	1.0058	24.7845**
9	MEANS	.0027	.0085	.1033	.6404	1.0055	22.1753**
10	PROPN	.0104	.0089	1.3730	.6427	1.0047	20.1346**
11	RELPR	-.0061	.0078	.6195	.6438	1.0052	18.3450**
12	NODP	-.0087	.0121	.5209	.6447	1.0060	16.8357**
13	MEANW	-.0032	.0086	.1387	.6450	1.0075	15.5119**
14	OTHP	-.0055	.0145	.1461	.6452	1.0091	14.3656**
15	SDS	.0027	.0085	.1033	.6453	1.0107	13.3719**
16	PREP	.0012	.0066	.0328	.6454	1.0124	12.4951**

N=297.
Intercept: 3.1650

* Significant at .05 level
** Significant at .01 level

TABLE 4.60
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- MIDDLE GROUP
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Multi-R	SE	F-value
1	LENGT	-.0522	.0063	69.2464**	.5603	.8132	119.4504**
2	KUYC	-.0167	.0072	5.4158*	.5887	.7952	68.9558**
3	MOST	-.0268	.0078	11.8236**	.6247	.7686	55.2741**
4	MOHC	-.0157	.0078	4.0708*	.6347	.7631	43.5172**
5	SPELL	.0081	.0050	2.6859	.6421	.7585	36.0670**
6	COMMA	-.0103	.0053	3.8082	.6489	.7543	31.0341**
7	MEANS	.0084	.0058	2.0523	.6539	.7515	27.2202**
8	OTHP	.0114	.0086	1.7550	.6583	.7492	24.2871**
9	HELPR	.0048	.0054	.7843	.6609	.7484	21.8020**
10	SDS	.0060	.0055	1.1852	.6626	.7483	19.7302**
11	PREP	.0036	.0053	.4482	.6632	.7493	17.9256**
12	SDW	-.0080	.0074	1.1845	.6638	.7503	16.4160**
13	DASH	.0060	.0064	.8855	.6653	.7505	15.2115**
14	MEANW	.0041	.0071	.3275	.6658	.7515	14.1099**
15	KEYP	-.0032	.0072	.1927	.6661	.7527	13.1385**
16	PROPN	.0020	.0059	.1129	.6663	.7541	12.2801**

N=263.
Intercept: 2.8973

* Significant at .05 level
** Significant at .01 level

ficity group in which the sentence unit was employed was slightly higher (.6663), but very close to the resulting MULTR for the corresponding analysis using the essay unit (Table 4.57).

Table 4.61 indicates the results obtained when the sentence was used as the search unit for content variable values in an analysis of the low content specificity group in which all sixteen predictors were employed. The significant variables which emerged were the same as those disclosed when the corresponding analysis for this group was carried out using the essay as search unit (Table 4.58), with one exception: the OTHP variable was not significant here. Another difference from the earlier analysis was that the MULTR obtained here, .5771, was somewhat lower than in the analysis employing the essay unit (.6002).

Using the content specificity level groups, a set of analyses were performed to determine the predictive level attained when the five content variables alone were used. The results of the three analyses using content variables only, where the essay was employed as the search unit for matches on the variables, are presented in Table 4.62. It is readily observed that best prediction resulted for the high content specificity group, with the difference in the MULTR values between the middle and low groups being negligible. This table also indicates a continuum in predictive efficiency for the various predictors moving from the high group to

TABLE 4.61
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- LOW GROUP
STYLISTIC AND CONTENT VARIABLES
(SENTENCE AS UNIT)

Step	Variable Entered	Final Step b-weight	SE	F-value	Cumulative MULT-R	SE	F-value
1	LENGT	-.0312	.0064	23.8331**	.4060	.9045	45.5984**
2	MODP	-.0222	.0073	9.3468**	.4870	.8663	35.7566**
3	MODC	-.0197	.0065	9.1987**	.5353	.8393	30.7389**
4	RELPR	-.0163	.0061	7.0548**	.5536	.8296	25.1956**
5	MEANS	.0055	.0064	2.2226	.5634	.8248	21.1232**
6	MEANW	.0030	.0064	.2280	.5680	.8235	17.9465**
7	KEYP	-.0086	.0080	1.1707	.5704	.8236	15.5073**
8	KEYC	-.0075	.0079	.8929	.5727	.8239	13.6681**
9	COMMA	.0042	.0062	.4631	.5738	.8249	12.1646**
10	DASH	.0042	.0064	.4414	.5748	.8261	10.9586**
11	SPELL	.0033	.0059	.3123	.5756	.8274	9.9577**
12	PROPN	.0030	.0064	.2280	.5762	.8288	9.1150**
13	OTHP	.0057	.0126	.2058	.5767	.8304	8.3948**
14	PREP	.0026	.0063	.1635	.5770	.8320	7.7749**
15	SDW	-.0013	.0079	.0267	.5771	.8339	7.2262**
16	SDS	-.0007	.0065	.0117	.5771	.8358	6.7445**

** Significant at .01 level

N=233.
Intercept: 3.1738

TABLE 4.52
STEPWISE MULTIPLE REGRESSION
CONTENT SPECIFICITY -- ALL GROUPS
CONTENT VARIABLES ONLY
(ESSAY AS UNIT)

Step	Variable Entered	Final Step			Cumulative		
		b-weight	SE	F-value	Mult-R	SE	F-value
High Group							
1	KEYP	-.0744	.0093	63.6239**	.4833	1.1305	80.9280**
2	KEYC	-.0529	.0097	30.0508**	.5991	1.0556	82.3421**
3	MODC	-.0242	.0127	3.6243	.6058	1.0309	56.6280**
4	OTHP	.0021	.0121	.0291	.6058	1.0336	42.3377**
5	MODP	-.0002	.0120	.0002	.6058	1.0370	33.7543**
N=227.		Intercept: 3.1630					
Middle Group							
1	KEYP	-.0281	.0074	14.4726**	.3284	.9287	30.7036**
2	KEYC	-.0263	.0084	9.7201**	.4675	.8696	36.3684**
3	MODP	-.0294	.0087	11.3388**	.4919	.8581	27.5690**
4	MODC	-.0225	.0091	6.0845**	.5096	.8496	22.6353**
5	OTHP	.0151	.0126	1.4355	.5135	.8489	18.4250**
N=268.		Intercept: 2.8973					
Low Group							
1	MODC	-.0324	.0071	32.0716**	.3259	.9357	27.4582**
2	MODP	-.0323	.0074	19.0601**	.4731	.8738	33.1644**
3	OTHP	.0338	.0104	10.6217**	.5057	.8575	26.2415**
4	KEYC	-.0146	.0085	2.0095	.5137	.8547	20.4399**
5	KEYP	-.0061	.0088	.4922	.5133	.8556	16.4139**
N=233.		Intercept: 3.1738					

* Significant at .05 level
** Significant at .01 level

the low group: for the high content specificity group, the teacher key-related variables are the significant predictors, whereas in the low group, the model student essay-related predictors and the variable representing matches on phrases from sources other than the teacher key and the model student essay are the significant variables. In the middle group, both teacher key and model essay-related variables are significant predictors. In the case of the high content specificity group, the MULTR obtained using the content variables alone was higher than the resulting value when the stylistic predictors alone were used in the equation (Table 4.53).

The use of the sentence as the search unit for matches on the content variables in a set of analyses using content variables only (Table 4.63) produced results very similar to those reported for the analysis described immediately above. Apparent differences between these last two sets of analyses were as follows: (1) the MULTR's obtained with the sentence as search unit were all somewhat lower than those achieved for the corresponding analysis when the essay was used as the search unit -- these differences increasing from high group to low; (2) the variable, "model student essay phrases completely matched," is significant for the high group only when the sentence is the search unit; (3) the

TABLE 4.63

STEPWISE MULTIPLE REGRESSION

CONTENT SPECIFICITY -- ALL GROUPS

CONTENT VARIABLES ONLY

(SENTENCE AS UNIT)

Step	Variable Entered	Final Step b-weight	Final Step SE	Final Step F-value	Cumulative Mult-R	Cumulative SE	Cumulative F-value
High Group							
1	KEYP	-.0475	.0093	52.5938**	.4721	1.1383	84.6118**
2	KEYC	-.0555	.0098	32.1421**	.5866	1.0476	77.1329**
3	MODC	-.0324	.0139	5.4215*	.5950	1.0405	53.7831**
4	MODP	-.0112	.0121	.8547	.5978	1.0398	40.5229**
5	OTHP	-.0041	.0139	.0883	.5979	1.0424	32.3548**
N=207.		Intercept: 3.1650					
Middle Group							
1	KEYP	-.0227	.0075	9.0522**	.3278	.9276	31.4304**
2	KEYC	-.0190	.0081	5.5586*	.4193	.8930	27.7406**
3	MODP	-.0302	.0088	11.6630**	.4498	.8803	31.9047**
4	MODC	-.0243	.0088	7.7309**	.4750	.8690	18.7975**
5	OTHP	-.0007	.0096	.0060	.4750	.8707	14.9813**
N=203.		Intercept: 2.8073					
Low Group							
1	MODP	-.0346	.0070	24.1438**	.3910	.9109	41.7073**
2	MODC	-.0208	.0066	9.8238**	.4569	.8822	30.3557**
3	KEYC	-.0124	.0080	2.3728	.4640	.8805	20.9527**
4	KEYP	-.0089	.0083	1.1534	.4681	.8803	16.0014**
5	OTHP	.0032	.0126	.0631	.4684	.8821	12.7612**
N=235.		Intercept: 3.1736					

* Significant at .05 level

** Significant at .01 level

variable, "phrases from other sources partially matched," is significant only for the low group and only when the essay is used as the search unit. Regarding items 2 and 3 above, the tendency, when the sentences replaces the essay as the search unit, for the model student essay-related variables to be more prominent as effective predictors and for the OTHP variable not to be found significant has been observed in previous analysis results (contrast the following pairs of tables: 4.46-4.49, 4.56-4.59, 4.57-4.60, 4.58-4.61).

A summary of the analyses by content specificity level is presented in Table 4.64. The increase in the MULTR value when both content and stylistic variables were used, rather than stylistic variables only, was greatest for the high and low groups. The middle content specificity level group, which had the highest original MULTR value in the "stylistic only" analysis made the smallest prediction gain when both content and stylistic variables were employed; nevertheless, the MULTR's attained when all sixteen predictors were used (whether with essay or sentence as search unit) were highest for the middle group. The independent variables were least efficient in predicting grades for the low content specificity group. When all sixteen independent variables were employed in the multiple regression equation, the

TABLE 4.64
SUMMARY OF ANALYSES BY CONTENT SPECIFICITY LEVEL

Group	Mult-R Stylis. Only	Mult-R Stylis./ Content (Essay Unit)	F_1	Mult-R Stylis./ Content (Sentence Unit)	F_2	Mult-R Content Only (Essay Unit)	F_3	Mult-R Content Only (Sentence Unit)	F_4
High (N=297)									
Orig.	.5163	.6477		.6454		.6058		.5986	
Shrunk.	(1) .4909**	(2) .6233**	14.0476**	(3) .6208**	13.7142**	(4) .5986**	1.2857	(5) .5902**	1.7142
Atten.	.5379	.6830	(1) vs. (2)	.6803	(1) vs. (3)	.6560	(2) vs. (4)	.6467	(3) vs. (5)
Middle (N=263)									
Orig.	.6009	.6565		.6663		.5136		.4750	
Shrunk.	(6) .5794**	(7) .6296**	5.0416**	(8) .6405**	6.4782**	(9) .5023**	5.4583**	(10) .4621**	6.9166**
Atten.	.6349	.6899	(6) vs. (7)	.7019	(6) vs. (8)	.5504	(7) vs. (9)	.5084	(8) vs. (10)
Low (N=233)									
Orig.	.4877	.6002		.5771		.5153		.4684	
Shrunk.	(11) .4512**	(12) .5621**	7.2258**	(13) .5356**	5.0303**	(14) .5025**	1.8387*	(15) .4535**	3.2258**
Atten.	.4944	.6160	(11) vs. (12)	.5869	(11) vs. (13)	.5506	(12) vs. (14)	.4969	(13) vs. (15)

* Significant at .05 level
** Significant at .01 level

resulting MULTR's were very similar regardless of whether the essay or sentence was used as the search unit for the content variable values. All comparisons made between the "stylistic only" MULTR's and those in which both content and stylistic variables were used resulted in highly significant $F_{difference}$ values. The multiple correlation coefficients obtained when only the five content variables were utilized in the equation were higher than the corresponding "stylistic only" analyses for the high and low content specificity groups but not for the middle group. MULTR's resulting from the inclusion of only the five content variables in the equation were compared with the corresponding full analysis (all sixteen predictors) for the same level with the same search unit. The $F_{difference}$ values representing these comparisons indicate that there was no significant difference in predictive ability for the high group when only the five content variables (rather than all sixteen predictors) were employed. However, for the middle and low groups, the use of all sixteen predictors yielded a significantly higher MULTR than did the use of only the five content variables. The middle content specificity group, in addition to achieving the best prediction results with the full variable set, is also unique among the three levels in yielding a lower MULTR when content variables only were used than the corresponding analysis for that level achieved when stylistic variables only were employed.

Summary

In this chapter, all descriptive statistics obtained on the essays under study were presented, followed by descriptions of the results of four multiple regression analysis series. Each such series was distinguished by the way in which the essays were grouped for analysis. In all four series, prediction coefficients for each group of essays were obtained initially using only the eleven stylistic predictors, then with the full complement of sixteen independent variables, both content and stylistic, first with the student essay as the search unit for determining values on the content variables and then with the essay sentence as the search unit for these values. In this way, three prediction coefficients were produced for each grouping level.

The four analysis series were:

- entire essay group
- grouping by subject area
- grouping by grade level
- grouping by content specificity level

In the first analysis series, where the entire group of essays were analyzed as one group, an empirical cross-validation was carried out to determine estimates of MULTR shrinkage. In all subsequent analyses, shrinkage estimates were obtained mathe-

matically using McNemar's equation. In all cases, the shrunken MULTR's were corrected for attenuation. Both the shrinkage and attenuation estimates were presented, with the corresponding original prediction coefficients, in summary tables following the description of each analysis series.

Two sets of analyses were completed with the essays in addition to those just described. In the first of these analyses, all essays were treated as one group, and the five content variables were combined to form two new content variables, "computer key phrases partially matched" (KEYP + MODP + OTHP) and "computer key phrases completely matched" (KEYC + MODC). These two "combined" content variables were then used, with the eleven stylistic predictors, in the multiple regression equation to determine what effect this new combination of the content variables would have on the resulting prediction level. As shown in Table 4.24, the MULTR's obtained with this set of predictors were comparable to those achieved when the five content variables were entered separately into the prediction equation.

Finally, a set of analyses were done with the three content specificity level groups in which only the five content variables were used in the multiple regression equation. The results of these analyses, displayed in Tables 4.62 and 4.63 and summarized

in Table 4.64, indicate that prediction was limited by the exclusion of the eleven stylistic variables from the equation, particularly for the middle and low content specificity groups, and most particularly when the sentence was used as the search unit for matches on the content variables.

Interpretation and discussion of all results is provided in Chapter V.

CHAPTER V

DISCUSSION AND IMPLICATIONS FOR FUTURE RESEARCH

The final chapter is divided into three major sections, the first presenting a discussion of the results obtained in the study, the second containing a summary of the hypothesis-testing results, and the third outlining some suggestions for future research in the area.

Discussion

Judge Reliability

The paired judge intercorrelations for the study ranged from .398 to .817, with seventeen out of the 24 coefficients being above .50. In general, the highest intercorrelations were found among the judges of the foreign language area essays, and the lowest values among the judges in the English area. Adjusted pooled reliability estimates obtained when the scores of all four judges in each subject area were used ranged from .785 (English) to .889 (foreign language). These results stand up well when compared to the judge reliabilities obtained in the studies cited in Chapter II -- even in comparison with a study such as Greene's (1970), in which rather elaborate procedures had been developed

and employed to insure judge reliability. The present reliability results, indeed, are surprisingly high in view of the fact that minimal precautions were taken to increase judge reliability. These precautions took the following forms: (1) teacher keys and exemplars (model student essays) were provided to the judges, and (2) the judges were requested to use the entire range of five possible scores in rating the set of essay responses related to each essay question.

Analyses of All Essays

The addition of the content variables to the multiple regression equation increased the MULTR value about .10 over the value obtained when the stylistic variables were used alone. This was true regardless of (1) which search unit was employed -- essay or sentence, and (2) whether the five content variables were entered into the equation individually or first combined into two variables (phrases completely matched and phrases partially matched). The multiple correlation coefficients obtained in this analysis set were somewhat lower than the values observed in corresponding analyses for more specific data groupings, especially those analyses in which both content and stylistic predictors were employed. An empirical cross-validation carried out on this data revealed some MULTR shrinkage; however, the results indicate that the

MULTR's were fairly stable.

Analyses by Subject Area

When the stylistic variables only were used in the multiple regression equations, the resulting MULTR's for the four subject areas could be divided into two clear groups on the basis of their size. The MULTR's for the science and social studies essay groups are very similar (.6010 and .6145, respectively), and the coefficients obtained for the English and foreign language essays were very close (.5017 and .4988). The length of essay variable was the most effective predictor for all subject areas except the foreign language area essay group (where it was not even a significant predictor). The significance of the proportion of spelling errors was unique to the science essay group; the foreign language essay group was distinguished by the prominence of the standard deviation of word length as an effective predictor. The proportion of proper nouns shows up, not surprisingly, as a significant predictor variable for the social studies and English essays groups; however, in both cases this is true only when certain combinations of predictors were used in the multiple regression equation. The proportion of relative pronouns was an important predictor only for the English essay group.

The inclusion of the five content variables in the prediction equation changed the rank orders of the MULTR sizes considerably.

Most dramatically, the foreign language area essay group, which had produced the lowest MULTR when stylistic variables only were used, showed the greatest gain in predictability when the content variables were added -- a difference of .21 between the MULTR when all sixteen variables were employed (regardless of which search unit was used) and the "stylistic only" MULTR. In fact, this group produced, with the inclusion of the five content variables in the prediction equation, the highest MULTR's found among the four subject area groups (.7099 when the essay was employed as the search unit, and .7120 when the sentence was the search unit for the content variables). The smallest gain in predictive value when the content variables were included in the regression equation was found for the social studies essay group (a difference of .07 when the essay search unit was used, and a difference of .05 when the sentence was used as the search unit).

Analyses by Grade Level

The length of essay variable was a highly significant predictor throughout all the analyses by grade level, regardless of the combination of variables employed in the prediction equation. It was the only significant variable which emerged in the "stylistic only" analysis of the 7th-8th grade level group. In the analyses of the 9th-10th grade essays, another stylistic variable

-- the proportion of spelling errors -- was an effective predictor, with essay length, for all three variable combinations employed in the analysis; this was the only grade level for which this variable was significant. The proportion of dashes showed up as a significant predictor in the 11th-12th grade "stylistic only" analysis, but did not appear prominently in the other two analyses conducted with this group. Two other stylistic variables were significant only when the sentence search unit was used: (1) the number of commas, for the 7th-8th grade level group, and (2) the proportion of relative pronouns, for the 11th-12th grade level group.

In general, the teacher key-related content variables -- especially the teacher key phrases partially matched, which was significant at every level, regardless of the search unit employed -- were the most effective predictors among the content variables. The teacher key-related variables were most prominent as predictors for the 11th-12th grade group of essays. A comparison of the two analyses for the 9th-10th grade level essay group in which all sixteen variables were employed reveals a phenomenon which was noted earlier in the analyses by subject area (see Tables 4.33 vs. 4.37), namely, that the use of the sentence as search unit tended to increase the predictove importance of the model student essay-related variables.

Best prediction results were obtained with the 9th-10th grade level group of essays, both when stylistic variables alone were used, and when all sixteen predictors were employed. The final coefficients obtained for the 7th-8th grade group and the 11th-12th grade group were very similar for all variable combinations tried. The greatest differences between "stylistic only" MULTR's and the MULTR's for the corresponding two analyses in which all sixteen predictors were employed occurred with the 11th-12th grade level group (a difference of about .14 in both cases), but these difference values are not much greater than those achieved for the 7th-8th grade level group.

Analyses by Content Specificity Level

Overall, the highest MULTR values for the usual three combinations of variables (stylistic only, stylistic and content variables with the essay as search unit, and stylistic and content variables with the sentence as search unit) were obtained for the middle content specificity group. The greatest gains in MULTR values when the content variables were included in the equations, however, were obtained for the high and low content specificity groups. When content variables alone were employed in the equations, the highest MULTR values were observed for the high content specificity group, the multiple correlation coefficients for the middle

and low content specificity groups being very similar. The most striking feature of the results obtained in this analysis set was the differential contribution of the various content variables at different content specificity levels. The teacher key-related variables were most prominent among the content predictors in the high content specificity level group. In the low group, the model student essay-related predictors and "other source-related" variable only, among the content variables, were significant predictors. In the results for the middle content specificity group, both teacher key and model student essay-related variables were represented among the significant predictors.

The significance of different stylistic predictors (other than the length of essay variable) at each content specificity level was another interesting feature of these results. The standard deviation of word length appeared as a significant predictor in all analyses of the high content specificity group in which it was used, while the proportion of relative pronouns was the only prominent stylistic predictor besides essay length in the analyses of the low content specificity group. The number of spelling errors appeared as a significant predictor in the stylistic analysis of the middle content specificity group only.

Overall Success of the Independent Variables in Prediction

The present study differs from earlier ones cited in its extensive exploration of the contributions made by both style-related and content-related variables to the prediction of essay grade. Table 5.1 presents a display of the performance of the various independent variables under the different types of variable groupings and data grouping arrangements used. The columns of the table indicate the particular combination of variables employed in an analysis. For the purposes of this presentation, the essay groupings employed in the study have been organized into eleven categories, coded as follows:

1 - All Essays
2 - Science Essays
3 - Social Studies Essays
4 - English Essays
5 - Foreign Language Essays
6 - Seventh-Eighth Grade Essays
7 - Ninth-Tenth Grade Essays
8 - Eleventh-Twelfth Grade Essays
9 - High Content Specificity Group Essays
10 - Middle Content Specificity Group Essays
11 - Low Content Specificity Group Essays

The cell entries are the essay grouping categories in the analyses of which a particular independent variable was found to be significant. This table, and the four tables following it, reveal the following:

-- The length of essay predictor was by far the most prominent variable in the analysis results, being a signifi-

TABLE 5.1
SIGNIFICANT PREDICTORS
FOR EACH PREDICTOR GROUPING AND DATA GROUPING COMBINATION

Variable	Style Only	Style and Content (Essay)	Style and Content (Sentence)	Style and Combined Content
LENGT	1,2,3,4,6,7, 8,9,10,11	1,2,3,4,6,7, 8,9,10,11	1,2,3,4,6,7, 8,9,10,11	1
SPELL	1,2,7,10	1,2,7	1,2,7	1
PREP				
PROPN	3,4	4		
MEANW	1,9			
SDW	1,5,9	1,5,9	1,5,9	1
MEANS				
SDS				
COMMA	9		1,2,6	1
DASH	8			
RELPR	4,11	4,11	4,8,11	
CONTC				1
CONTP				1

			Content Only
KEYC	**	1,5,7,8,9,10	1,3,5,8,9,10	9,10
KEYP		1,2,3,5,6,7,8,9	1,2,5,6,7,8,9	9,10
MODC		1,2,4,6,11	1,2,4,6,7,9,10,11	9[a],10,11
MODP		1,2,3,4,10,11	1,2,4,7,10,11	10,11
OTHP		3,4,6,11		11

[a]A significant predictor in this analysis only when the sentence was used as the search unit.

Note: Identified suppressor variables not included in this table.

TABLE 5.2
SIGNIFICANT PREDICTORS
ANALYSES USING ALL ESSAYS

Style Only	Style and Content (Essay)	Style and Content (Sentence)	Style and Combined Content (Essay)	Style and Combined Content (Sentence)
(4.17)[a]	(4.18)	(4.19)	(4.20)	(4.21)
LENGT	LENGT	LENGT	LENGT	LENGT
MEANS	KEYP	KEYP	CONTP	CONTP
SPELL	KEYC	KEYC	CONTC	CONTC
SDW	SDW	MODP	SDW	SDW
	MODC	MODC	SPELL	SPELL
	MODP	SDW	COMMA	COMMA
	SPELL	SPELL		
		COMMA		

[a]Numbers of the tables displaying the results from which the information in each cell was obtained are given in parentheses.

TABLE 5.3
SIGNIFICANT PREDICTORS
ANALYSES BY SUBJECT AREA

Subject Area	Style Only	Style and Content (Essay)	Style and Content (Sentence)
Science	(4.29)[a] LENGT SPELL	(4.33) LENGT MODC KEYP SPELL MODP	(4.37) LENGT MODP MODC COMMA KEYP SPELL
Social Studies	(4.30) LENGT PROPN	(4.34) LENGT KEYP MODP OTHP	(4.38) LENGT KEYC
English	(4.31) LENGT RELPR PROPN	(4.35) MODC RELPR OTHP MODP LENGT PROPN	(4.39) MODP MODC LENGT RELPR
Foreign Language	(4.32) SDW	(4.36) KEYP KEYC SDW	(4.40) KEYP KEYC SDW

[a]Numbers of the tables displaying the results from which the information in each cell was obtained are given in parentneses.

TABLE 5.4
SIGNIFICANT PREDICTORS
ANALYSES BY GRADE LEVEL

Grade Level	Style Only	Style and Content (Essay)	Style and Content (Sentence)
7th-8th	(4.42)[a]	(4.45)	(4.48)
	LENGT	KEYP	LENGT
		LENGT	KEYP
		MODC	MODC
		OTHP	COMMA
9th-10th	(4.43)	(4.46)	(4.49)
	LENGT	LENGT	LENGT
	SPELL	KEYP	MODP
		KEYC	MODC
		SPELL	SPELL
			KEYP
11th-12th	(4.44)	(4.47)	(4.50)
	LENGT	LENGT	LENGT
	DASH	KEYC	KEYC
		KEYP	KEYP
			RELPR

[a]Numbers of the tables displaying the results from which the information in each cell was obtained are given in parentheses.

TABLE 5.5
SIGNIFICANT PREDICTORS
ANALYSES BY CONTENT SPECIFICITY LEVEL

Group	Style Only	Style and Content (Essay)	Style and Content (Sentence)	Content Only (Essay)	Content Only (Sentence)
High	(4.53) LENGT MEAMW COMMA SDW	(4.56) KEYP KEYC LENGT SDW	(4.59) KEYP KEYC LENGT SDW MODC	(4.62) KEYP KEYC	(4.63) KEYP KEYC MODC
Middle	(4.54) LENGT SPELL	(4.57) LENGT KEYC MODC	(4.60) LENGT KEYC MODP MODC	(4.62) KEYP KEYC MODP MODC	(4.63) KEYP KEYC MODP MODC
Low	(4.55) LENGT RELPR	(4.58) LENGT MODC MODP OTHP RELPR	(4.61) LENGT MODP MODC RELPR	(4.62) MODC MODP OTHP	(4.63) MODP MODC

[a]Numbers of the tables displaying the results from which the information in each cell was obtained are given in parentheses.

cant predictor in all but one of the eleven types of analyses (the foreign language area essay set).

-- Of the ten remaining stylistic predictors, only four were significant predictors in analyses in which the content variables also were employed. These were: (1) the proportion of spelling errors, (2) standard deviation of word length, (3) the number of commas, and (4) the proportion of relative pronouns.

-- An interesting phenomenon in the results was the appearance of the number of commas as a significant predictor in combination with the content variables only in those analyses in which the sentence search unit was used.

-- In those analyses in which the content variables were employed, at least two content variables emerged as significant predictors in each case, regardless of the two search units used.

-- The "other source-related" content variable was the least effective content predictor, and was a significant predictor only in those analyses in which the essay was used as the search unit.

-- Of the other four content variables, the variable representing model student essay phrases completely matched

was most influenced by the type of search unit employed. The other three content variables were little affected by the shift from essay search unit to sentence search unit.

Major Hypothesis-Testing Results

The major hypothesis underlying the study was that the inclusion of variables derived from consideration of both content and style in the multiple regression analyses of student essays would prove significantly more effective in predicting student essay grades than would the use of stylistic variables alone. The 24 comparisons made in this study between MULTR's obtained using stylistic variables only and those obtained when a combination of stylistic and content variables were employed all support this hypothesis -- the differences between the corresponding MULTR pairs yielding a significant F-value in all cases.

Regarding the four subject areas, it had been hypothesized that gain in the MULTR values from the addition of the content variables to the multiple regression equations would be greater for the two groups of essays (science and foreign language essays) whose questions appeared likely to require more specific responses than did the items in the other two subject areas (English and social studies). This hypothesis was not upheld. When the content variables were employed in combination with the stylistic

variables, the greatest gains in MULTR values were observed in those subject areas which had clearly produced the lowest multiple correlation coefficients when stylistic variables only were used (English and foreign language). All of the $F_{difference}$ values for the analyses by subject area were highly significant. It is possible that the greater gains noted for the English and social studies areas may in part be artefacts of the measurement procedure -- i.e., that a "ceiling effect" determining the highest MULTR value attainable with the present data (partially due to imperfect reliability of the criterion) may have served to depress the gains made by those groups whose initial (stylistic only) MULTR's were relatively high.

A third hypothesis was that the MULTR values resulting from the addition of content variables to the equation would be higher for the 7th-8th grade level group than for the 11th-12th grade level group, under the assumption that in the lower grade level, content, more than stylistic, considerations would be emphasized by the judges. Here also the observed results did not support this hypothesis. For all three combinations of variables employed, the highest MULTR's were produced by the 9th-10th grade level essays. However, the greatest gains in MULTR values occurring with the addition of the content variables were observed for the

7th-8th grade level groups, with very little difference noted in the corresponding MULTR values for the two groups over the various predictor combinations. As in the results associated with the previous hypothesis, the major factor influencing the amount of gain made by the MULTR in these cases seems to have been the initial value of the MULTR produced using the stylistic variables only.

The fourth hypothesis underlying the study concerned the content specificity groupings of the essays, holding that the gain in MULTR values resulting from the addition of the content variables would be greater for the high content specificity group than for the low content specificity group. This turned out to be indeed the case. While the highest multiple correlation coefficients for the usual three combinations of predictors occurred using the middle content specificity group, the least gain in MULTR value was observed for this group. The high content specificity group made the greatest gain in predictability with the addition of the content variables, regardless of whether the sentence or the essay was employed as the search unit for the content variables. The particularly great contribution of the content variables in predicting the criterion measures for the high content specificity group is better revealed by an examination of the MULTR values obtained by the three groups using content

variables only. The MULTR's in this case are highest for the high content specificity group and, in fact, the two MULTR's for this group (using the two different search units) are not significantly different from the corresponding values obtained when both the content and stylistic variables were employed. The MULTR values obtained using content predictors only are very similar for the middle and low content specificity groups, and these values are considerably lower than the values for the high content specificity group.

The final hypothesis was that the choice of search unit for the content variables would not have any significant effect on the resulting MULTR values. The seventeen comparisons which were made in connection with this hypothesis all support it.

Implications for Future Research

Computer Algorithms

As the reader may have noted from the discussion in Chapter III, the present study does not take full advantage of the variety of options available in the computer programs which were prepared for the AEC-II Project. The algorithms which provide part-of-speech information on the words in the text under analysis are of particular interest. Using this information, linguistic analyses might be performed which could provide insight into

the syntactical structures preferred by students of differing grade levels and in different essay examination situations. Another application of this approach might be for the identification of syntactical features which present special problems to students. These are only a few suggestions from a wide variety of potential research topics in this area.

The prepared programs are limited in their ability to analyze textual data in that the following important considerations are not reflected in the content analysis algorithms:

(1) procedures for handling anaphora (since the antecedents may be key content words).

(2) procedures for special handling of passive-voice verbs: even though a particular response unit may be found to contain all words present in a key phrase, the present scoring procedure, which (a) allows for any permutation of word order within the response unit, and (b) reduces all content words to root-word form, with resulting loss of inflectional information, makes it possible for the student to receive full credit for a phrase match even in those instances in which, by the use of a passive voice verb or by the insertion of a word or phrase of negation, he has stated the complete opposite of what was called for by the essay question.

The effectiveness of the content-related predictors produced in the absence of such considerations in this study, however, is quite encouraging.

The "Thesaurus Problem"

The absence of a manageable, general, computer-usable thesaurus and associated algorithms for its use continues to be a major problem in the area of essay analysis, as it was in 1969 when Marcotte mentioned this drawback in connection with his discussion of needed research in this area. This problem was not so much of a hindrance in the studies done by McManus and Marcotte, where the type of essay under scrutiny was the response to a "short answer" essay item, designed to require in response brief, content-laden, sentences or phrases containing specific words and dates. Indeed, McManus, dealing with such data, was able to employ successfully an empirical item discrimination technique in which a computer program produced statistics which determined the "key words" to be used for the content analysis of the essays. Here, too, the specific nature of the essay response required of the student enabled McManus to disregard completely possible synonyms for the relevant content words disclosed in the empirical selection process. The "thesaurus problem," however, becomes critical as the complexity and/or "openness" of the essay response called for increases.

The time-consuming nature and the excessive tedium of the thesaurus-generating process can be attested to by anyone who has embarked on a content analytic study, of however limited a range. This task was easily the most time-consuming one engaged in during the present study. Hopefully, however, the close scrutiny of the textual data base necessitated by the manual thesaurus-generating process for content analytic tasks will contribute insights into the nature of natural language data which may collectively form the basis for a manageable computer-usable thesaurus and appropriate algorithms for its use. The collection of such information forms one fruitful area for immediate inquiry.

Content Specificity

Based on the results of this study, further exploration of the content specificity factor in essay rating would appear promising. One reason for the attraction of this area of investigation is the apparent relationship between content specificity and inter-judge agreement (in relation to this, see Marcotte, 1969, p. 70), and analysis of this relationship may suggest methods for enhancing judge reliability -- an important consideration in studies of the type under investigation here. The current data provide tentative clues for distinguishing potentially valuable predictors for essay items requiring high content-specific responses versus those which

do not. The lack of significant difference in the MULTR obtained for the high specificity group when content variables only were used versus the use of the full complement of predictors indicates the negligible contribution of stylistic variables to prediction for this group of essays. These findings coincide with McManus' results (1968, p. 57) for short answer history items at the college level, in which the inclusion of two stylistic predictors (number of commas and standard deviation of word length) with the content-related variables in the multiple regression equation did not, in general, enhance the MULTR values appreciably. An optimal scheme for prediction of essay grade with such groups, suggested by the existing data, would seem to be to employ previously-identified useful content predictors in conjunction with a "relevancy variable" of the type used by McManus (in all of his analyses, this variable was the best single predictor of essay grade -- p. 57).

It would seem that the content specificity dimension could have been better defined than it was in the present study. High content specificity need not preclude an emphasis on stylistic considerations also, for example. Future studies exploring this dimension should perhaps avoid a definition of content specificity which is made in terms of the absence of stylistic emphasis.

Content-Related Predictors

The particular content variables selected for the present investigation were chosen from a large number of possible predictors. General statements regarding the efficacy of these specific predictors must await the results of future studies comparing their contributions to prediction with those of other content-related variables, using the same type of data. The development of well-defined set of effective predictors, and the identification of the situations in which different content variables work best as predictors, is one large area for further exploration in textual analysis.

This study did not attempt to replicate the method employed in Marcotte's study, in which three-sentence clusters in the student essays were used as response units, in addition to the single-sentence student response mode. Here, instead, the two student response units investigated were whole-essay and single-sentence. According to Marcotte's results, the three-sentence scoring unit for the content-related variables produced values more highly correlated with the criterion measures than did the single-sentence scoring unit. Since Marcotte's sample consisted of student responses to college-level short-answer examinations, it remains to be seen whether his "loose scoring" technique would enhance prediction results with data from a younger population, over the results obtained using either whole-essay or single-sentence as response

units.

Restricted Models

The large number of comparisons necessary to complete the minimum requirements of the present study precluded the inclusion of even a brief discussion of work which has been begun on an investigation of potentially effective restricted models which might be employed with the present data, including those which make use of interaction terms. These results remain to be developed in a later paper.

Summary

In summary, the current study has been an attempt to provide initial data on the analysis of essays covering: (1) both content and stylistic predictors of the criterion, essay grade; (2) a number of subject areas; (3) several grade levels; (4) several item content specificity levels; and (5) two "response modes" or search units for values on the content variables -- whole-essay and single-sentence. The study yields evidence for the following:

-- the differential roles of the individual predictors in various levels of the data groupings

-- the predictive importance of the length of essay

-- the ability of the chosen content variables to enhance prediction and, associated with this, the implied efficacy of the

procedures chosen for generating these variables.

It is apparent from much of the foregoing discussion that, while certain findings of previous studies have been replicated here, a number of new areas of research on this topic have been made visible, requiring extensive further exploration.

BIBLIOGRAPHY

Agne, Russell. "Data Collection." Appendix 8 in Page and Longyear, Analysis of Essays by Computer. Progress Report, 1969.

Aiken, L.R., Jr. "The Grading Behavior of a College Faculty," Educational and Psychological Measurement, 1963, 23 (2), 319-322.

Anderson, C.C. "The New STEP Test as a Measure of Composition Ability," Educational and Psychological Measurement, 1960, 20, 95-102.

Archambault, Francis X. A Computerized Approach to Scoring Verbal Responses to the Torrance Tests of Creative Thinking. Unpublished doctoral dissertation, the University of Connecticut, 1969.

_____. A Computerized Approach to the Detection of Spelling Errors. Unpublished masters thesis, the University of Connecticut, 1968.

Baldwin, A.L. "Personal Structure Analysis: A Statistical Method for Investigating the Single Personality," Journal of Abnormal and Social Psychology, 1942, 37, 163-183.

Berelson, B. Content Analysis in Communication Research. Glencoe, Ill.: The Free Press, 1952.

Bloom, B.S. "Learning for Mastery." Evaluation comment. UCLA Center for Study in Instructional Programs, 1968, 1 (2), 12 pp.

Bomberger, R. An Analytical Comparison of Scope-Depth Balance with Popularity of General Circulation Weekly Magazines in the United States. Cited by Budd, R., and Thorp, R., eds., An Introduction to Content Analysis. Iowa City, Iowa: University of Iowa School of Journalism, 1963.

Box, G.E.P., and Cox, D.R. "An Analysis of Transformations," Journal of the Royal Statistical Society, Series B, 1964, 26, 211-252.

Braddock, Richard; Lloyd-Jones, Richard; and Schoer, Lowell. Research in Written Composition. Champaign, Ill.: National Council of Teachers of English, 1963.

Budd, R., and Thorp, R. An Introduction to Content Analysis. Iowa City, Iowa: University of Iowa School of Journalism, 1963.

Carroll, J.B. "A Factor Analysis of Verbal Ability," Psychometrika, 1941, 6, 279-307.

_____. The Study of Language: A Survey of Linguistics and Related Disciplines in America. Cambridge, Mass.: Harvard University Press, 1953.

Coffman, William E. "Essay Examinations". Chapter 10 in R.L. Thorndike, ed., Educational Measurement. 2nd ed. Washington, D.C.: American Council on Education, 1971, pp. 271-302.

_____. "On the Validity of Essay Tests of Achievement," Journal of Educational Measurement, 1966, 3, 151-156.

_____, and Kurfman, D.A. "A Comparison of Two Methods of Reading Essay Examinations," American Educational Research Journal, 1968, 5, 99-107.

College Entrance Examination Board. Annual Report of the College Board 1965-1966. New York: College Entrance Examination Board, 1966.

Coward, A.F. "The Method of Reading the Foreign Service Examination in English Composition," Educational Testing Service Research Bulletin #57, 1950.

Daigon, A. "Computer Grading of English Composition," The English Journal, January 1966, 46-52.

Diederich, P. B. "Cooperative Preparation and Rating of Essay Tests," English Journal, 1967, 56, 573-584.

_____. "Definitions of Ratings on the ETS Composition Scale." Princeton, N.J.: Educational Testing Service, 1962. (Multilith)

_____; French, J.W.; and Carlton, S.T. "Factors in the Judgments of Writing Ability." ETS Research Bulletin, #61-15. Princeton, N.J.: Educational Testing Service, 1961.

_____, and Link, F.R. "Cooperative Evaluation in English." In F.T. Wilhelms, ed., Evaluation as Feedback and Guide. Washington, D.C.: Association for Supervision and Curriculum Development, NEA, 1967, pp. 181-232.

DuBois, P.H. "A Test-Dominated Society: China, 1115 B.C.-1905 A.D." In Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1965, pp. 3-11.

Duke University Computation Center. TSAR (Tele-Storage and Retrieval System) User's Manual. Duke University, September 1970.

Dunphy, D.C.; Stone, P.J.; and Smith, M.S. "The General Inquirer: Further Developments in a Computer System for Content Analysis of Verbal Data in the Social Sciences," Behavioral Science, 1965, 10, 468-480.

Dutch, Robert A., ed. The Original Roget's Thesaurus of English Words and Phrases. New York: Dell Publishing Co., Inc., 1962.

Ebel, R.L. "Estimations of the Reliability of Ratings," Psychometrika, 1951, 16, 407-424.

_____. Measuring Educational Achievement. New York: Prentice-Hall, 1965.

Educational Testing Service. "Judges Disagree on Qualities That Characterize Good Writing," ETS Developments, 9 (2), February 1961.

Efroymson, M.A. "Multiple Regression Analysis." In Ralston and Wilf, eds., Mathematical Methods for Digital Computers. New York: John Wiley and Sons, 1960.

Finlayson, D.S. "The Reliability of the Marking of Essays," British Journal of Educational Psychology, 1951, 21, 126-134.

Fisher, G. "The SCORTXT Program for the Analysis of Natural Language." Storrs, Conn.: The University of Connecticut Bureau of Educational Research, 1968. (Mimeograph)

Fosvedt, D.R. "Criteria for the Evaluation of English Composition," Journal of Educational Research, 1965, 3, 108-112.

Fox, Karl A., and Ezekiel, Mordecai. Methods of Correlation and Regression Analysis, Linear and Curvilinear. New York: John Wiley and Sons, 1959.

Francis, W.N. The Structure of American English. New York: The Ronald Press Co., 1958.

French, J.W. "The Effects of Essay Tests on Student Motivation." ETS Research Bulletin #4. Princeton, N.J.: Educational Testing Service, 1956.

French, J.S. "Schools of Thought in Judging Excellence of English Themes." In Proceedings of the 1964 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1962, pp. 19-28.

Garber H. "Computer Grading Versus the Human Judge." Paper read at the Annual Meeting of the American Psychological Association, Chicago, Illinois, September 7, 1965.

Garvin, P.L., and Spolsky B., eds. Computation in Linguistics. Bloomington, Ind.: Indiana University Press, 1966.

George, A. "Quantitative and Qualitative Approaches to Content Analysis." In I. Pool, ed., Trends in Content Analysis. Urbana, Ill.: University of Illinois Press, 1959.

Gleser, G.C.; Cronbach, L.J.; and Rajaratnam, N. "Generalizability of Scores Influenced by Multiple Sources of Variance," Psychometrika, 1965, 30, 395-418.

Godshalk, F.I.; Swineford, F.; and Coffman, W.E. The Measurement of Writing Ability. New York: The College Entrance Examination Board, 1966.

Gosling, G.W.H. Marking English Compositions. Victoria: Australian Council for Educational Research, 1966.

Grant, D.L., and Caplan, N. "Studies in the Reliability of the Short-Answer Essay Examination," Journal of Educational Research, 1957, 51 (2), 109-116.

Greene, John F. Scoring Creativity Tests by Computer Simulation. Unpublished doctoral dissertation, The University of Connecticut, 1970.

Guilford, J.P.; Christensen, P.R.; Taafe, G.; and Wilson, R.C. "Ratings Should be Scrutinized," Educational and Psychological Measurement, 1962, 22 (3), 439-447.

Gustav, A. "Students' Preferences for Test Format in Relation To Their Test Scores," Journal of Psychology, 1964, 57, 159-164.

Hartog, P., and Rhodes, E.C. The Marks of Examiners. New York: Macmillan, 1936.

Hayakawa, S.I. "Learning to Think and Write: Semantics in Freshman English," Journal of the Conference on College Composition and Communication, February 1962.

Hays, David G. Introduction to Computational Linguistics. New York: American Elsevier, 1967.

Hayes, R.M. "Mathematical Models in Information Retrieval." In P.L. Garvin, ed., Natural Language and the Computer. New York: McGraw-Hill Book Co., Inc., 1963, pp. 268-309.

Hiller, J.H.; Fisher, G.; and Kaess, W. "A Computer Investigation of Verbal Characteristics of Effective Classroom Learning," American Educational Research Journal, 1969 6 (4), 661-675.

_____; Marcotte, D.R.; and Martin, T. "Opinionation, Vagueness, and Specificity Distinctions," American Educational Research Journal, 1969, 6 (2), 271-286.

Hoffman, B. The Tyranny of Testing. New York: Collier Books, 1962.

Holsti, O. "Content Analysis Research in the Social Sciences. Paper read at IBM Texas A&M Conference on Computers in Humanistic Research, College Station, Texas, November 1966.

Hunt, K.W. "Differences in Grammatical Structures Written at Three Grade Levels. USOE Cooperative Research Project 1998. Tallahassee, Fla.: Florida State University, 1964.

Iker, H.P., and Harway, N.I. "A Computer Approach Towards the Analysis of Content," Behavioral Science, 1965, 10, 173-182.

Janzen, H.L. A Study of Written Language Ability. Unpublished masters thesis, The University of Calgary, Alberta, Canada, 1968.

Jewett, A., and Bish, C.E., eds. Improving English Composition. Washington, D.C.: National Education Association, 1965.

Lesk, M.E. "A Proposal for English Text Analysis." Bell Telephone Laboratories, 1969.

Marcotte, D.R.; Page, E.B.; and Daigon, A. "A Computer Analysis of Cliché Behavior in Student Prose." Paper read at the Annual Meeting of the American Educational Research Association, New York, February 18, 1967.

_____. A Computerized Contingency Analysis of Content Graded Essays. Unpublished doctoral dissertation, The University of Connecticut, 1969.

Marshall, J.C. "Composition Errors and Essay Examination Grades Reexamined," American Educational Research Journal, 1967, 4, 375-386.

McColly, W., and Remstad, R. "Comparative Effectiveness of Composition Skills Learning Activities in the Secondary School." USOE Cooperative Research Project 1528. Madison, Wisc.: University of Wisconsin, 1963.

McManus, John F. Computer Evaluation of College History Examinations by Actuarial Strategies. Unpublished doctoral dissertation, The University of Connecticut, 1968.

McNemar, Q. Psychological Statistics. 3rd ed. New York: Wiley, 1962.

Merkhofer, B.E. College Students' Study Behavior. Unpublished doctoral dissertation, The University of Chicago, 1950.

Meyer, G. "An Experimental Study of the Old and New Types of Examination: I. The Effect of the Examination Set on Memory," Journal of Educational Psychology, 1934, 25, 641-661.

_____. "An Experimental Study of the Old and New Types of Examination: II. Methods of Study," Journal of Educational Psychology, 1935, 25, 641-661.

_____. "The Choice of Questions on Essay Examinations," Journal of Educational Psychology, 1939, 30, 161-171.

Morehead, Albert H., and The National Lexicographic Board, eds. The New American Roget's College Thesaurus in Dictionary Form. New York: Grosset and Dunlap, 1958.

Myers, A.E.; McConville, C.B.; and Coffman, W.E. "Simplex Structure in the Grading of Essay Tests," Educational and Psychological Measurement, 1966, 26, 41-54.

Noyes, E.S. "Essay and Objective Tests in English," College Board Review, 1963, 49, 7-10.

Osgood, C.E.; and Anderson, L. "Certain Relations Among Experienced Contingencies, Associative Structure, and Contingencies in Encoded Messages," American Journal of Psychology, 1957, 70, 411-420.

Page, E.B. "Analysis of Essays by Computer (AEC-II)." Research Proposal, March 1969.

_____. "The Imminence of Grading Essays by Computer," Phi Delta Kappan, 1966, 47, 238-243.

_____. "Grading Essays by Computer: Progress Report." In Proceedings of the 1966 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1967, pp. 87-100.

_____, and Paulus, D.H. The Analysis of Essays by Computer. Final Report. Project No. 6-1318, Contract No. OEC-10-001318-1214. U.S. Department of Health, Education, and Welfare. Storrs, Conn.: The University of Connecticut, 1968.

Page, E.B. "Statistical and Linguistic Strategies in the Computer Grading of Essays." 2ieme Conference Internationale Sur Le Traitement Automatique Des Langues (Proceedings), August 23-25, 1967, No. 34.

Pearson, R. "The Test Fails as an Entrance Examination." In "Should the General Composition Test Be Continued?" College Board Review, 1955, No. 25, 2-9.

Scannell, D.P., and Marshall, J.C. "The Effect of Selected Composition Errors on Grades Assigned to Essay Examinations," American Educational Research Journal, 1966, 3, 125-130.

Schultz, L. "Language and the Computer." In H. Borko, ed., Automated Language Processing. New York: John Wiley and Sons, Inc., 1967, pp. 11-31.

Schuster, D.H. "The Quasi-Essay Test, a New Type of Test." Paper read at the Annual Meeting of the American Psychological Association, Washington, D.C., September 1967.

Sedelow, S.Y., and Sedelow, W.A., Jr. "Stylistic Analysis." In H. Borko, ed., Automated Language Processing. New York: John Wiley and Sons, Inc., 1967, pp. 181-213.

Serling, A.M. "Three Methods of Testing Literary Comprehension at the Advanced Placement Level." College Board Research and Development Reports 66-7, No. 4. Princeton, N.J.: Educational Testing Service, 1967.

Simmons, R.F. "SYNTHEX: Toward Computer Synthesis of Human Language Behavior." In H. Borko, ed., Computer Applications in the Behavioral Sciences. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1962, pp. 360-393.

Slotnick, H.B. "Toward a Theory of Essay Grading by Computer." National Assessment of Educational Progress, 1971.

_____, and Knapp, J.V. "Essay Grading by Computer: A Laboratory Phenomenon?" English Journal, 1971, 60, 75-87.

Solomon, R.J. "Improving the Essay Test in the Social Studies," National Council on Social Studies Yearbook, 1965, 35, 137-153.

Stalnaker, J.M. The Essay Type of Examination." In E.F. Lindquist, ed., Educational Measurement. Washington, D.C.: American Council on Education, 1951, pp. 495-530.

Stanley, J.C. "Analysis of Variance Principles Applied to the Grading of Essay Tests," Journal of Experimental Education, 1962, 30, 279-283.

Starch, D., and Elliott, E.C. "Reliability of Grading of High School Work in English," School Review, 1912, 20, 442-457.

Stone, P.J.; Bales, R.F.; Namenwirth, J.Z.; and Ogilvie, D.M. "The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information," Behavioral Science, 1962, 7, 484-498.

_____; Dunphy, D.; Smith, M.; and Ogilvie, D. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, Mass.: MIT Press, 1966.

Storey, A.G. "Some evidence in the Essay-Item Case," The Journal of Educational Research, 1968, 61 (8), 351-354.

Swineford, F. "Test Analysis, Advanced Placement Tests, Form EBP." Statistical Report No. 28. Princeton, N.J.: Educational Testing Service, 1956.

_____. "Test Analysis, Advanced Placement Examination in American History, Form MBP." ETS Statistical Report No. 53. Princeton, N.J.: Educational Testing Service, 1964.

_____. "Test Analysis, Advanced Placement Examination in Mathematics, Form MBP." ETS Statistical Report No. 52. Princeton, N.J.: Educational Testing Service, 1964.

Taylor, C.W. "A Factorial Study of Fluency in Writing," Psychometrika, 1947, 12, 239-262.

Tillett, P. "Documentation of Programs Prepared for the AEC-II Project." Storrs, Conn.: The University of Connecticut, 1972. (Xerox)

Vallance, T.R. "A Comparison of Essay and Objective Examinations as Learning Experiences," Journal of Educational Research, 1947, 41, 279-288.

Vernon, P.E., and Millican, G.D. "A Further Study of the Reliability of English Essays," British Journal of Statistical Psychology, 1954, 7, 65-74.

Whalen, T.E. A Computer Analysis of Mechanical Proficiency and Overall Quality in Seventh Grade Essays. Unpublished doctoral dissertation, The University of Connecticut, 1970.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Co., Inc., 1962.

Winograd, Terry. Understanding Natural Language. New York: Academic Press, 1972.

Wiseman, S. "The Marking of English Composition in Grammar School Selection," British Journal of Educational Psychology, 1949, 19, 200-209.

Wiseman, S., and Wrigley, S. "Essay Reliability: The Effect of Choice of Essay Title," Educational and Psychological Measurement, 1958, 18, 129-138.

Wittenborn, J.R. "Manual Ability, Its Nature and Measurement," Educational and Psychological Measurement, 1945, 5, 241-260; 395-409.

Wood, D.A. Test Construction. Columbus, Ohio: Charles Merrill Books, Inc., 1961.

Young, D. "Examining Essays for Eleven Plus Classification," British Journal of Educational Psychology, 1962, 32, 267-274.

Zicky, Michael J. An Automated Aid to the Study of Literature: The Application of Content Analytic and Statistical Techniques to Paradise Lost. Unpublished doctoral dissertation, The University of Connecticut, 1968.

APPENDIX A

INSTRUCTIONS TO ESSAY RATERS

OCTOBER 1971 and MARCH 1972

Expertise

"You have been chosen as an essay grader because you are a competent teacher in this general field of study. You don't have to be an expert in the specific subareas for your judgments to be very helpful."

The Essays

"There are twelve sets of answers/essays to be graded. These have been obtained in response to questions given to students ranging from seventh grade to twelfth grade, during ordinary class periods. The questions were set by the teachers of the classes on subject matter studied in class.

"Six classes are represented. For each class there are two different questions and hence two sets of essays.

"The order in which the twelve sets are presented to each grader has been randomized; the order of the essays within each set has also been randomized. The essays and associated material are presented in the form of a computer printout."

Practice Set and Procedure

"You have been given a practice set of essays; these essays will not be used in the grading session proper, but are provided to familiarize you with the procedure to be followed.

"Refer to the first page of the practice set. At the beginning of each set appears identifying information. Opposite the word, "Session," the date and time of commencement of grading should be written. This is followed by the question for the set. The next paragraph identifies the class from which the essays were obtained. This is followed by a key to grading, provided by the teacher. (In some sets this key may be missing.)

"You are asked to assess the usefulness of the key before commencing to grade the student essays. You will also be asked to do this after completing the set.

"The next item is a "model" essay, regarded by the teacher as a good response. (In some sets this model may be missing.) You are asked to assess the usefulness of the model both before and after grading the set. (Refer to the last page of the practice set.)

"Then follows the student essays to be graded. Preceding each essay and to the left of it is a rating scale, immediately above which are figures which will be used subsequently to identify each essay and are of no significance in grading."

Rating the Essays

"In rating the essays, it is essential that you use the full range of ratings in each set of essays. There should be at least one essay in each set with rating "1" (high), and at least one with rating "5" (low), and all ratings from 1 to 5 should be used in each set. This should be done even though you may consider, for example, that in a particular set there is no essay worthy of rating 1 by your normal standards. In other words, we require relative ratings, from 1 to 5, among the essays of each set, not gradings based on absolute standards. Because of this requirement, you may find it advisable to read the first five or so essays in each set before commencing to assign ratings for that set, in order to ascertain the general standard of essays in the set.

"In some sets there are essays completely blank (i.e., the student supplied no response). You are asked to assign a rating to these "essays" also.

"The ratings should be based on evidence of the student's knowledge of the subject matter, but the grade may be influenced by style or expression if you regard such matters as important in your own teaching practice. You are not obliged to adhere to the criteria set forth in the teacher's key, or implied by the model student essay."

Limitations of the Computer Printout

"As you will have noticed, the use of computer printout means that all letters are capitalized. Apart from this, the essays have been reproduced as submitted by the students. You may assume that spelling errors, incomplete sentences, etc., present, occurred in the original essay. However, note that a dash "on the line," i.e., level with the bottom of the letters, denotes that the word it precedes was underlined in the original essay."

Questions

"Questions should be asked freely during the practice session, to avoid interruptions as far as possible in subsequent sessions. Start the practice set."

APPENDIX B

LIST OF FUNCTION WORDS USED AS INPUT TO STAGE 3, RUN 1

a	everyone	off	thou	within
about	everything	on	though	without
across	for	onto	thus	you
after	from	or	thy	your
along	he	our	to	yourself
although	her	ourselves	toward	
among	herself	overall	towards	
an	him	re	unless	
and	himself	she	until	
as	his	so	upon	
at	how	some	us	
because	however	such	we	
beneath	I	than	what	
but	if	that	whatever	
by	in	the	when	
despite	inside	thee	whenever	
during	into	their	where	
each	it	theirs	whereas	
either	its	them	wherever	
else	itself	themselves	whether	
enough	me	these	which	
ever	neither	they	who	
every	no	thine	whom	
everybody	nor	this	why	
everyman	of	those	with	

APPENDIX C

THE STEMMING ALGORITHM
(STAGE 3, RUN 3)

The Stemming Algorithm

The suffix removals currently provided for by the program (SUBROUTINE STEMER) are shown in the following table:

Any suffix removal is applied only if all the following are true:

(1) The length of the full word is not less than the figure in column 1.
(2) The corrected spelling does in fact end with the suffix indicated in column 2.
(3) The stemmed spelling, obtained by removing the suffix in column 2 and adding the suffix in column 3, is in the dictionary.
(4) The dictionary entry has the syntax shown in column 4, unmodified, as one of its entries.
(5) The dictionary entry yields the same root word as that previously obtained for the current word, or, no root word has previously been obtained for this spelling.
(6) There are less than four syntax outputs already assigned to this word.
(7) The proposed syntax output, shown in column 5, does not already exist in the syntax already assigned.

When all the above conditions are met, the syntax shown in column 5 is added to the syntax area in the first available blank space, the stemmed spelling is placed in the STEM output array, if blank, and the dictionary root word is placed in the root word array DOUT if blank.

All stemmings appropriate to the spelling of the current word are examined, regardless of the success or failure of previous attempts at analysis, e.g., removal of ST, EST, and DEST would all be attempted for SADDEST (even if the full spelling SADDEST occurred in the dictionary).

Many apparent omissions from the suffix set (e.g., the formation of adjectives by adding AL) are not so because such inflections are regularly included in the Merriam-Webster Pocket Dictionary (of which the tape dictionary employed in this run may be regarded as a subset). This is almost treu of some suffixes which have been included, e.g., the formation of adverbs

1	2	3	4	5	EXAMPLE
4	D		VB	VBT2	HATED
4	'D		PN	VBT2	THEY'D
5	ED		VB	VBT2	ALLOWED
5	IED	Y	VB	VBT2	MARRIED
5	'RE		PN	VB	THEY'RE
5	ING		VB	VBP1	ALLOWING
5	ING	E	VB	VBP1	LOVING
4	'LL		PN	VB	THEY'LL
5	MEN	MAN	N	N PL	BUSINESSMEN
4	R		AJ	AJCO	FINER
4	R		VB	N	LOVER
5	ER		AJ	AJCO	SWIFTER
5	ER		VB	N	EMPLOYER
5	DER		AJ	AJCO	MADDER
5	DER		VB	N	PLODDER
5	IER	Y	AJ	AJCO	HAPPIER
5	IER	Y	VB	N	MULTIPLIER
4	S		N	N PL	THINGS
4	S		PR	PRPL	AMERICAS
4	S		VB	VBT1	HATES
4	'S		N	N PO	THING'S
4	'S		PN	VBT1	SHE'S
5	ES		N	N PL	DRESSES
5	ES		VB	VBT1	HISSES
5	IES	Y	N	N PL	FLIES
5	IES	Y	VB	VBT1	COMPLIES
6	N'T		VB	VB	HASN'T
6	ST		AJ	AJCO	FINEST
6	EST		AJ	AJCO	HIGHEST
6	DEST		AJ	AJCO	SADDEST
6	IEST	Y	AJ	AJCO	HAPPIEST
4	LY		AJ	AV	NORMALLY

by addition of LY to the adjective, and the formation of nouns by adding ER to the verb. These latter have been included to cater for certain "invented" words in the essays which do not occur in the MPD.

The algorithm is not recursive, which may be regarded as its chief defect. Words formed by the addition of two or more suffixes will not be recognized, e.g., DOESN'T = DO + ES + N'T. Very few such cases arise in the essay set, however, and these have been made special entries in the dictionary.

The suffixes have all been embedded in the coding of the algorithm. A superior approach would have been to write a general algorithm using the suffix data via table lookup, the table being created from card input to the program.

In certain instances, the suffix may indicate that the spelling is a combination of two words (THEY'D, THEY'RE, THEY'LL, SHE'S) of which the first is normally a pronoun. While the first may be a noun (THE PLAY'S THE THING), this case has not been provided for by the algorithm. In the above cases, the pronoun constituent has been dropped from the output, and the verb root form (WILL, HAVE, BE) put into the root word area by the algorithm (if the root word was previously blank). The program has no provision for replacing one input spelling by two separate output spellings.

APPENDIX E

INSTRUCTIONS TO RATERS

RATINGS OF ESSAY QUESTIONS AS TO CONTENT SPECIFICITY AND EMPHASIS

RATINGS OF CONTENT SPECIFICITY AND EMPHASIS

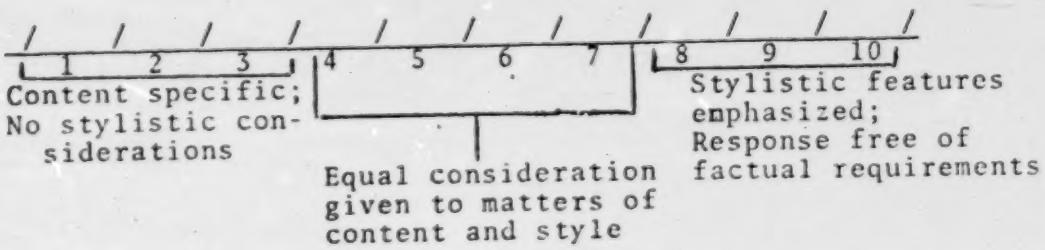
This package contains a series of typical short-answer or essay questions that have been used by teachers at the junior high school and high school level.

As a teacher, you are being asked to rate these questions on a dimension of content specificity and emphasis.

You should rate those questions which seem to you to require the student to present only specific facts and which do not place any emphasis on stylistic considerations toward the "1" end of the continuum. Those questions which you perceive to be so broad that no single list of specific facts or phrases could be used to represent the "ideal" response to the question should be placed toward the "10" end of the continuum (such questions would also be those for which stylistic features of the responses would be emphasized in scoring the responses).

Please remember that you are not being asked to rank order the questions. All questions in a given package may be judged as having "high" content specificity and emphasis or "low" content specificity and emphasis.

A diagram showing the general areas of the continuum is presented below. The continuum has been divided into ten parts in order to allow you to represent more precisely the degree of content specificity and emphasis you think is implied in a particular question.


1 2 3
4 5 6 7
8 9 10
Content specific; No stylistic considerations
Equal consideration given to matters of content and style
Stylistic features enphasized; Response free of factual requirements

Using these criteria, please rate the questions in the accompanying package.

APPENDIX E

INSTRUCTIONS TO RATERS

RATINGS OF ESSAY QUESTIONS AS TO CONTENT SPECIFICITY AND EMPHASIS

RATINGS OF CONTENT SPECIFICITY AND EMPHASIS

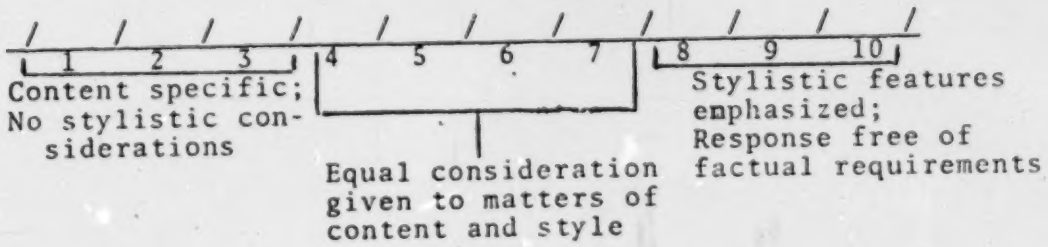
This package contains a series of typical short-answer or essay questions that have been used by teachers at the junior high school and high school level.

As a teacher, you are being asked to rate these questions on a dimension of content specificity and emphasis.

You should rate those questions which seem to you to require the student to present only specific facts and which do not place any emphasis on stylistic considerations toward the "1" end of the continuum. Those questions which you perceive to be so broad that no single list of specific facts or phrases could be used to represent the "ideal" response to the question should be placed toward the "10" end of the continuum (such questions would also be those for which stylistic features of the responses would be emphasized in scoring the responses).

Please remember that you are not being asked to rank order the questions. All questions in a given package may be judged as having "high" content specificity and emphasis or "low" content specificity and emphasis.

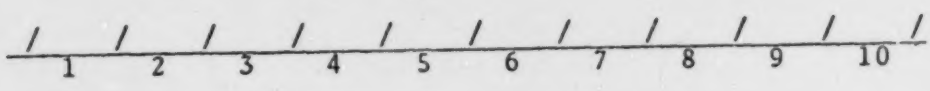
A diagram showing the general areas of the continuum is presented below. The continuum has been divided into ten parts in order to allow you to represent more precisely the degree of content specificity and emphasis you think is implied in a particular question.

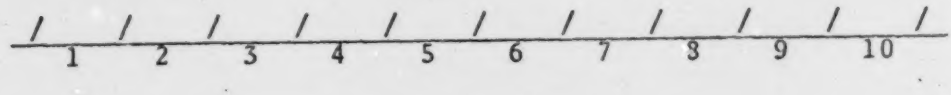

1 2 3
4 5 6 7
8 9 10
Content specific; No stylistic considerations
Equal consideration given to matters of content and style
Stylistic features emphasized; Response free of factual requirements

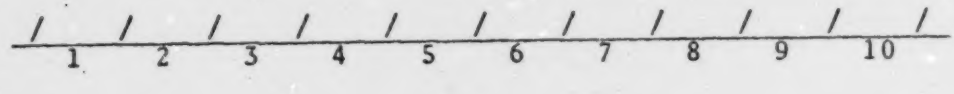
Using these criteria, please rate the questions in the accompanying package.

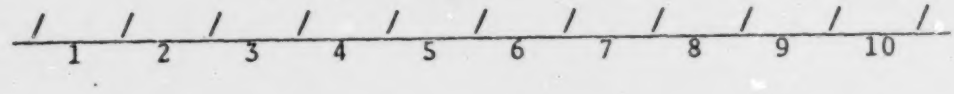
Subject Area: _____

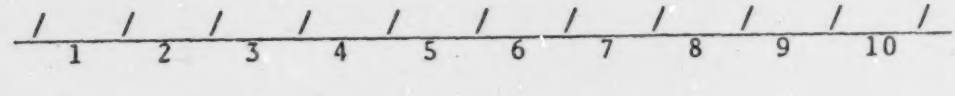
Rater # _____

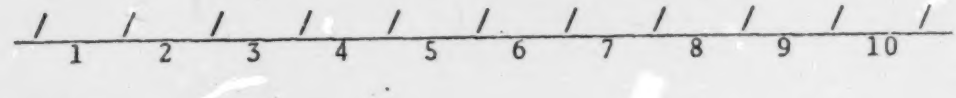
_____. 

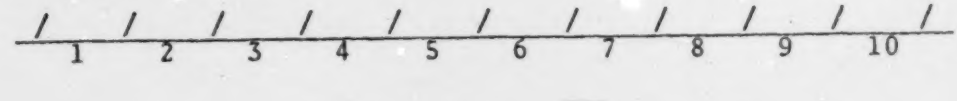
_____. 

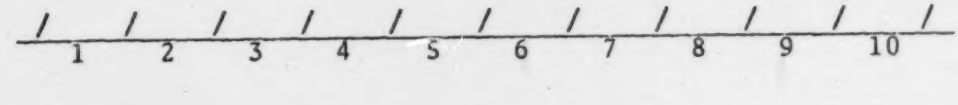
_____. 

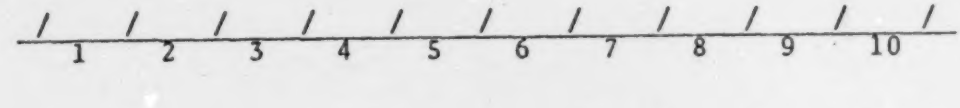
_____. 

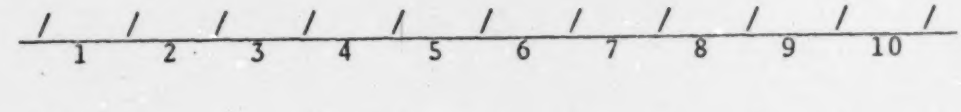
_____. 

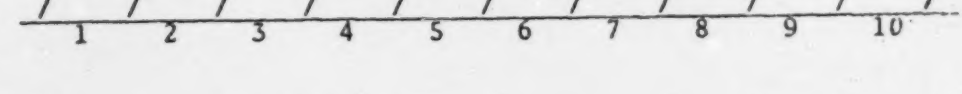
_____. 

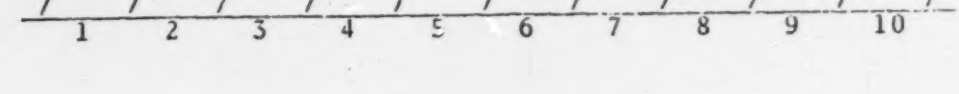
_____. 

_____. 

_____. 

_____. 

_____. 

_____. 

_____.